



BOTIKA

Big Data and Its Implementation

Hello...

Rahayu Oktavia Putri

Data Team at Botika

Email : <u>rahayuoktavia01.ro@gmail.com</u>

Telegram: R_oktavia

Whatsapp: 085736921381

Nilam Mufidah

Data Team at Botika

Email : <u>nilam19mufidah@gmail.com</u>

whatsapp : 085655599366



Agenda

- Introduction to Big Data
- Crawling Data from Twitter
- Data Preparation
- EDA
- Machine Learning



Introduction To Big Data



What Is Big Data?



OR



Is this big enough?

It should be like this?

The most people say that

Big Data is a big amount of data being collected



Hmmm.... What do you think about it?

But the fact is **not** just about **size or volume**.

Big Data refers to the large amounts of data that is pouring in from various data sources and has different formats, which is difficult to store and process using available database management tools or traditional data processing applications.





Talk About Format

Structured

Big Data

Semi-Structured

Unstructured

Any Idea?



Structured Data

	A	В	С	D	E	F	G		Н	l .		K	L M	N	0
	hotel	s_canceled lead	d_time a	arrival_date_year arriva	l_date_month	arrival_date_week_number a	rrival_date_day_of_mon	th sta	ys_in_weekend_nights stays_ir	n_week_nights adu	s chil	ldren k	pabies meal	country	market_segment
2	Resort Hotel	0	342	2015 July		27		1	0	0	2	0	0 BB	PRT	Direct
3	Resort Hotel	0	737	2015 July		27		1	0	0	2	0	0 BB	PRT	Direct
4	Resort Hotel	0	7	2015 July		27		1	0	1	1	0	0 BB	GBR	Direct
5	Resort Hotel	0	13	2015 July		27		1	0	1	1	0	0 BB	GBR	Corporate
6	Resort Hotel	0	14	2015 July		27		1	0	2	2	0	0 BB	GBR	Online TA
7	Resort Hotel	0	14	2015 July		27		1	0	2	2	0	0 BB	GBR	Online TA
8	Resort Hotel	0	0	2015 July		27		1	0	2	2	0	0 BB	PRT	Direct
9	Resort Hotel	0	9	2015 July		27		1	0	2	2	0	0 FB	PRT	Direct
10	Resort Hotel	1	85	2015 July		27		1	0	3	2	0	0 BB	PRT	Online TA
11	Resort Hotel	1	75	2015 July		27		1	0	3	2	0	0 HB	PRT	Offline TA/TO
	Resort Hotel	1	23	2015 July		27		1	0	4	2	0	0 BB	PRT	Online TA
13	Resort Hotel	0	35	2015 July		27		1	0	4	2	0	0 HB	PRT	Online TA
14	Resort Hotel	0	68	2015 July		27		1	0	4	2	0	0 BB	USA	Online TA
15	Resort Hotel	0	18	2015 July		27		1	0	4	2	1	0 HB	ESP	Online TA
16	Resort Hotel	0	37	2015 July		27		1	0	4	2	0	0 BB	PRT	Online TA
17	Resort Hotel	0	68	2015 July		27		1	0	4	2	0	0 BB	IRL	Online TA
18	Resort Hotel	0	37	2015 July		27		1	0	4	2	0	0 BB	PRT	Offline TA/TO
19	Resort Hotel	0	12	2015 July		27		1	0	1	2	0	0 BB	IRL	Online TA
20	Resort Hotel	0	0	2015 July		27		1	0	1	2	0	0 BB	FRA	Corporate
21	Resort Hotel	0	7	2015 July		27		1	0	4	2	0	0 BB	GBR	Direct
22	Resort Hotel	0	37	2015 July		27		1	1	4	1	0	0 BB	GBR	Online TA
	Resort Hotel	0	72	2015 July		27		1	2	4	2	0	0 BB	PRT	Direct
24	Resort Hotel	0	72	2015 July		27		1	2	4	2	0	0 BB	PRT	Direct
25	Resort Hotel	0	72	2015 July		27		1	2	4	2	0	0 BB	PRT	Direct
26	Resort Hotel	0	127	2015 July		27		1	2	5	2	0	0 HB	GBR	Offline TA/TO
	Resort Hotel	0	78	2015 July		27		1	2	5	2	0	0 BB	PRT	Offline TA/TO
28	Resort Hotel	0	48	2015 July		27		1	2	5	2	0	0 BB	IRL	Offline TA/TO
29	Resort Hotel	1	60	2015 July		27		1	2	5	2	0	0 BB	PRT	Online TA
30	Resort Hotel	0	77	2015 July		27		1	2	5	2	0	0 BB	PRT	Online TA
	Resort Hotel	0	99	2015 July		27		1	2	5	2	0	0 BB	PRT	Online TA
32	Resort Hotel	0	118	2015 July		27		1	4	10	1	0	0 BB	NULL	Direct
33	Resort Hotel	0	95	2015 July		27		1	4	11	2	0	0 BB	GBR	Offline TA/TO
34	Resort Hotel	1	96	2015 July		27		1	2	8	2	0	0 BB	PRT	Direct
	Resort Hotel	0	69	2015 July		27		2	2	4	2	0	0 BB	IRL	Offline TA/TO
36	Resort Hotel	1	45	2015 July		27		2	1	3	3	0	0 BB	PRT	Online TA
	Resort Hotel	1	40	2015 July		27		2	1	3	3	0	0 BB	PRT	Online TA
	Resort Hotel	0	15	2015 July		27		2	1	3	2	0	0 BB	ESP	Online TA
	Resort Hotel	0	36	2015 July		27		2	1	3	3	0	0 BB	PRT	Online TA
	Resort Hotel	1	43	2015 July		27		2	1	3	3	0	0 BB	PRT	Online TA
		-		Lozo odiy	(1)				-				700	print	

Semi-Structured data

```
("created at":"Sat May 26 13:18:31 +0000 2018","id":1000365564357832705,"id str":"1000365564357832705","text":"RT @madridismoreaI : Mi XI para ma\u00flana: Navas, Carvajal, Nacho, Ramos, Marcelo, Modric, Kroos, Casemiro, Isco
{"created at":"Sat May 26 13:18:31 +0000 2018","id":1000365566224486400,"id str":"1000365566224486400","text":"RT @LFC: #UCLfinal day. \ud83d\udd34\n\n#WeAreLiverpool https:\/\/t.co\/eJwkX0v7vV","source":"\u003ca href=\"http:
{"created at":"Sat May 26 13:18:31 +0000 2018","id":1000365567000350722,"id str":"1000365567000350722","text":"RT @LorisKarius: #UCLFINAL \nAllez Allez \ud83d\udd34\ud83d\ude4c\ud83c\udffc @LFC https:\/\/t.co\/x1x01scF1
{"created at":"Sat May 26 13:18:31 +0000 2018","id":1000365567402958848,"id str":"1000365567402958848","text":"RT @AyoDesigns: #UCLFinal.\nWho's winning it all? https:\/\/t.co\/jfz4pnF8wA","source":"\u003ca href=\"http:\/\/two
{"created at":"Sat May 26 13:18:32 +0000 2018","id":1000365569311375362,"id str":"1000365569311375362","text":"RT @goal: Excitement level: \ud83d\udcaf\n\nReal Madrid vs. Liverpool: In words and numbers. \ud83d\ude4c\n\n#UCLf
```

20150608_GrexitByCountry.csv	688 bytes	CSV docum	04 Februari 2020, 17:31
20150611_CopaAmericaTeam	256 bytes	CSV docum	04 Februari 2020, 17:31
20150611_MERSvsEbola.csv =	526 hytes	CS/\ docum	04 Februari 2020, 17:31
20150615_BushPostAnnounce	2,2 kB	CSV docum	04 Februari 2020, 17:31
20150615_BushPostAnnounce	1,7 kB	CSV docum	04 Februari 2020, 17:31
20150615_BushPreAnnounce	2,2 kB	CSV docum	04 Februari 2020, 17:31
20150615_BushPreAnnounce	1,7 kB	CSV docum	04 Februari 2020, 17:31
20150615_NBAFinalsGame5.csv	6,0 kB	CSV docum	04 Februari 2020, 17:31
20150616_2016 Candidates In R	5,4 kB	CSV docum	04 Februari 2020, 17:31
20150617_DenmarkElections.csv	22,4 kB	CSV docum	04 Februari 2020, 17:31
20150617_NBAFinalsPlayers.csv	684 bytes	CSV docum	04 Februari 2020, 17:31
20150623_ConfederateFlag.csv	2,7 kB	CSV docum	04 Februari 2020, 17:31
20150624_FathersDay.csv	12,7 kB	CSV docum	04 Februari 2020, 17.31
20150624_NBADraft.csv	2,1 kB	CSV docum	04 Februari 2020, 17:31
20150626_ElectionCandidates	7,1 kB	CSV docum	04 Februari 2020, 17:31
20150626_SameSexMarriage.csv	2,9 MB	CSV docum	04 Februari 2020, 17:31
20150629_WimbledonMenSin	2,6 kB	CSV docum	04 Februari 2020, 17:31
20150629_WimbledonWomen	2,9 kB	CSV docum	04 Februari 2020, 17:31
20150630_4julymovies.csv	13,4 kB	CSV docum	04 Februari 2020, 17:31
20150701_WWCvsCopaAmeri	1,3 kB	CSV docum	04 Februari 2020, 17:31
20150705_greecereferendum	2,1 kB	CSV docum	04 Februari 2020, 17:31
20150707_SandersVsClinton.csv	835 bytes	CSV docum	04 Februari 2020, 17:31
20150708_AshesPlayers.csv	687 bytes	CSV docum	04 Februari 2020, 17:31
20150709_SSEvsGREXIT.csv	11,7 kB	CSV docum	04 Februari 2020, 17:31
20150710_WimbledonMensSi	5,5 kB	CSV docum	04 Februari 2020, 17:31
20150710_WimbledonWomen	5,9 kB	CSV docum	04 Februari 2020, 17:31
20150713_ScottWalkerTopCiti	2,5 kB	CSV docum	04 Februari 2020, 17:31
20150713_ScottWalkerTopSta	3,6 kB	CSV docum	04 Februari 2020, 17:31

1	Search interest in Ebola and MERS by month globally, indexed by 100		
2	month	Search interest in Ebola	Search interest in MERS
3	January 2014	0	1.64
4	February 2014	0.08	1.54
5	March 2014	0.51	1.68
6	April 2014	1.05	19.53
7	May 2014	0.34	41.34
8	June 2014	0.42	6.28
9	July 2014	1.64	3.36
10	August 2014	13.65	2.76
11	September 2014	3.26	2.77
12	October 2014	13.28	4.73
13	November 2014	2.37	2.64
14	December 2014	1.06	1.77
15	January 2015	0.69	1.53
16	February 2015	0.53	4.08
17	March 2015	0.51	3.51
18	April 2015	0.31	1.81
19	May 2015	0.25	4.87
20	June 2015	0.29	100
21			

1	Search interest in Jeb Bush by city, following his 6/15 announcement	t .
2	city	Search Interest
3	Kendall (Florida United States)	100
4	Town and Country (Missouri United States)	94.03
5	Bala Cynwyd (Pennsylvania United States)	69.2
6	Coral Gables (Florida United States)	62.99
7	The Hammocks (Florida United States)	53.17
8	Fountainableau (Florida United States)	48.08
9	Cape Coral orida United States)	43.4
10	Palm Coast (Florida United States)	42.74
11	Weston (Florida United States)	40.71
12	Port St. Lucie (Florida United States)	36.24
13	Greenwich (Connecticut United States)	35.94
14	Palm Beach Gardens (Florida United States)	33.55
15	Hyattsville (Maryland United States)	33.45
16	Tallahassee (Florida United States)	33.41
17	Miami (Florida United States)	33.37
18	Davie (Florida United States)	32.06
19	Burke (Virginia United States)	30.49
20	Alexandria (Virginia United States)	29.02
21	Bradenton (Florida United States)	28.87
22	Williamsburg (Virginia United States)	28.68
23	West Palm Beach (Florida United States)	28.34
24	Yonkers (New York United States)	27.42
25	Lone Tree (Colorado United States)	25.97
26	Fort Lauderdale (Florida United States)	25.76
27	Doral (Florida United States)	25.58
28	Boca Raton (Florida United States)	24.86

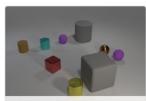
Unstructured Data



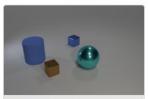
CLEVR_test_000000.png Size 164.6 KB



CLEVR_test_000001.png Size 170.07 KB



CLEVR_test_000002.png Size 188.94 KB



CLEVR_test_000003.png Size 181.11 KB



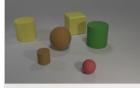
CLEVR_test_000004.png Size 188.03 KB



CLEVR_test_000005.png Size 166.13 KB



CLEVR_test_000006.png



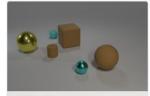
CLEVR_test_000007.png Size 183.39 KB



CLEVR_test_000008.png Size 179.96 KB



CLEVR_test_000009.png Size 189.78 KB



CLEVR_test_000010.png Size 185.81 KB



CLEVR_test_000011.png Size 177.17 KB



CLEVR_test_000012.png Size 189.85 KB



CLEVR_test_000013.png Size 182.73 KB



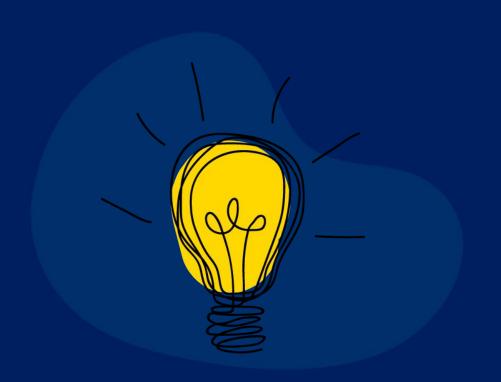
CLEVR_test_000014.png Size 180.34 KB



CLEVR_test_000015.png Size 175.42 KB

```
0-26.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.08615384615384615, "y":0.3063063063063
            0de0ee708a4a47039e441d488615ebb7.png", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.7053087757313109, "y
            0rCMbYS7vkSINlu61.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.46956521739130436,"y":0.05},{";
            _l.png","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.495,"y":0.27147766323024053},,{"x":0.595,"y":0.460481
            1f913f304d4265684fea960106d1132a when-families-live-together featuredImage.jpg.jpeg","annotation":[{"label":["Face"],"ı
            2a63b028915ebd600062f8a552ba097c.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.486,"y":0.046},
            2:867CA8F00000578-3199800-image-a-1_1439712734572.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0
           2.5.5 - 2.hd9pmj.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.46875,"y":0.27111111111111114},{"x":0.56
            _3c4f3c819d8771e425ba09462ce2148e--funny-man-so-funny.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{";
            3d---4K---2D-HD-DTS-ATMOS-4000--MOVIES---9632354635-FIRST-CONFIRM-VB201705171774173-ak WBP638920203-1524385027
           4.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.48, "y":0.10385756676557864}, {"x":0.771
            _4ac2f735-8419-4c89-910e-a9e6174f3017-1.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.0462962962
           05-families-fighting-for-change-vogue-february-2018.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x
            05HelpingFamiliesGetAhead 0.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.24268689057421453, "y
            5 job interview movie scenes.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.6389568052159739, "y
            _5a0bcde9a83fcd89048b464e-750-375.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.2626666
            5b2d5ccbla0000c504ce261f.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.218851570964247,"y
            5c8670b4917b6d7efe68c4f73042f96f.jpq.jpeq", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.306608
           6c9f5fafffdf652e5bbf672fd355e4de.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.421746293245469
           __07d4fe77dce1c871944c80c3f5596b5c--funny-movie-quotes-funny-movies.jpg.jpeg","annotation":[{"label":["Face"],"notes":"
            7c9e84ab6ald7cd45032c69f612893bc--the-train-cousin.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x"
            7e09d2b84-1.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.108333333333333, "y":0.180555555555
            08slid1.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.7904761904761904,"y":0.19948186528497408
            10-coverfield-lane-275x350.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.22545454545454546,"y"
            0013.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.49, "y":0.20285714285714285}, {"x":0.58, "y":0.
            13 Favorite Memorable Movie Scenes Misery.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.335,"y"
            014a8f16-c5ce-43fc-bf88-a4950cac2394.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.31,"y":0.134
            14-scenes-from-movies-and-tv-that-made-us-cry-in--2-28533-1513793846-10 dblbig.jpg.jpgg","annotation":[{"label":["Face
            lbbest-scenes14.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.6402985074626866,"y":0.056300268
            18q0dngvi7ic2jpg.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.3725,"y":0.2746268656716418},{":
            19c50e32afad4fb81558d57296645da8.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.1468531468531468
            19-E0Y-sex-scenes.nocrop.w710.h2147483647.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.0704225
            22-reunification-migrant-families.w710.h473.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.38732
            22-reunite-families.w710.h473.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.1619718309859155,"
                         e97b8437542d375794e.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.0875,"y":0.04429
            34 American Beauty.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.7638136511375948,"y":0.2192307
            57_dark_knight.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.4263888888888888,"v":0.3746530989
            065b36-20180619-activists-in-l-a-protest-separation-of-migrant-children-from-families.jpg.jpeg", "annotation":[{"label"
            067efdb73cb3b14502617091325bcb78365827bfb4dc6a8eedae87269a810331.jpg.jpeg","annotation":[{"label":["Face"],"notes":""
            80smovies.png", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.6996124031007752, "y":0.36176470588235293}
            82-5.jpg.jpeg", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.31716417910447764, "y":0.2885714285714286}
           94e2c06402f1b04c585fa9c1895d7967.png", "annotation":[{"label":["Face"], "notes":"", "points":[{"x":0.17979797979798, "y
           _825a26941df80f8fdb3569ed6c1ee999.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.295161290322580
9201d1901f 920x920.jpg.jpeg","annotation":[{"label":["Face"],"notes":"","points":[{"x":0.1608695652173913,"y":0.1104199066874028}
```

But why we should know about that? It is just a format right?



Define the tools that you need

Every format need a different way to analyze

Need a different method

Need different ways to be solved

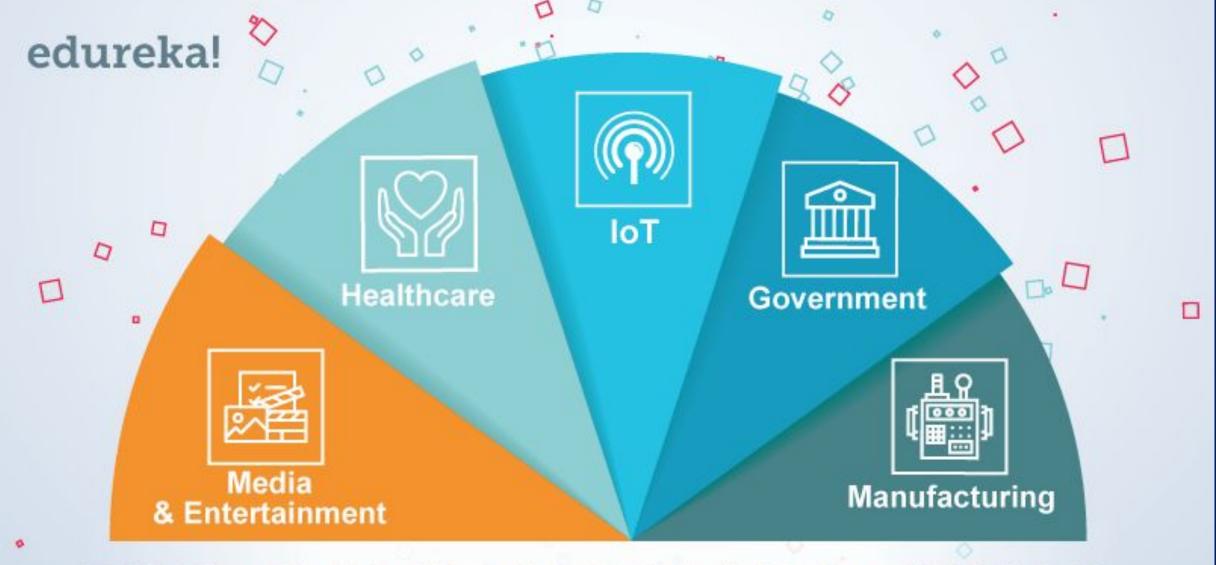
Characteristics of big data

Useful Data



and Tools used

Implementation of big data



BIG DATA APPLICATIONS

Crawling and Scraping Data



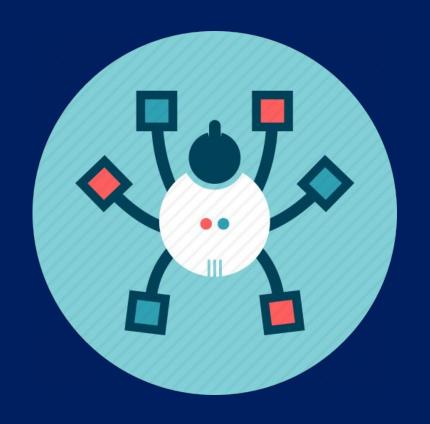


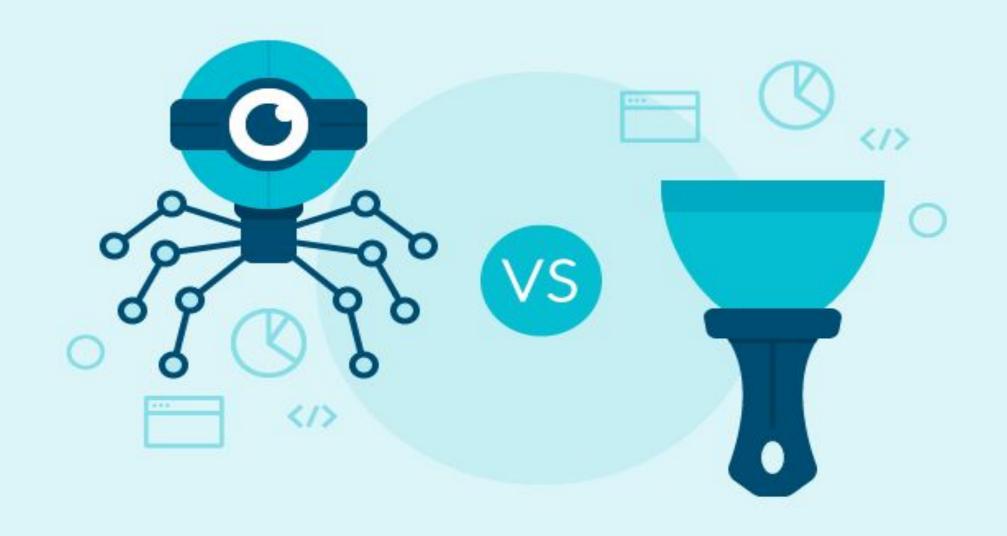
Scraping Data

Data scraping is the finding of data and then scraping it. it pulls data directly from a page. this doesn't pull exclusively from the web, it can be taken from anywhere that data exist.

Crawling Data

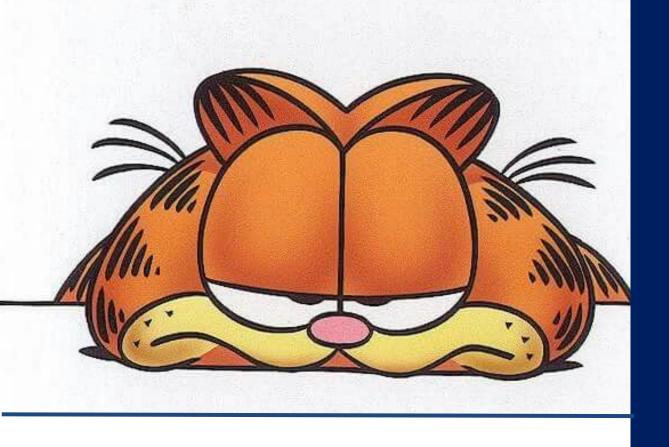
web crawling is digging deep into the nooks and crevices of the world wide web to retrieve the stuff you missed in spring cleaning. think about spiders or bots, scavenging through the web to find whatever is relevant to your quest





Data Preparation





I feel bored!!!!!

Spend almost 80% of our time

But

It is IMPORTANT

What wrong with our data?



- Have different pattern
- Dublicates
- Missing value
- There is outliers
- Unexpected data values
- Doesn't fit with our need

Dig a little deeper to your data

Knowing your data
Cleaning
Manipulating Data

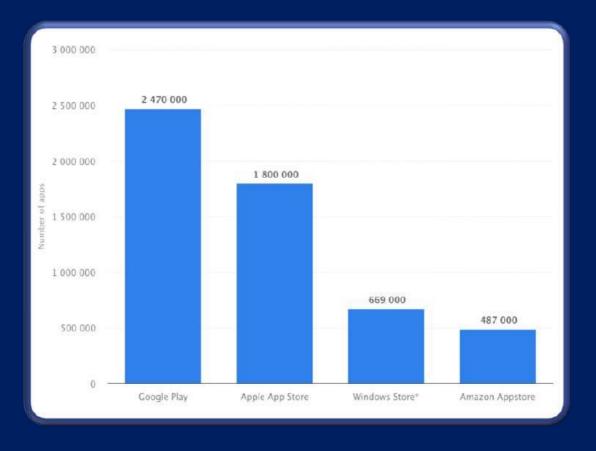
So, what should we do?



Exploratory Data Analysis (EDA)







Source:

https://www.statista.com/statistics/



You won't understand, if you don't try Let's check jupyter notebook file



Machine Learning





Type of Machine Learning Algorithms

Let's make it simple

Supervised Learning

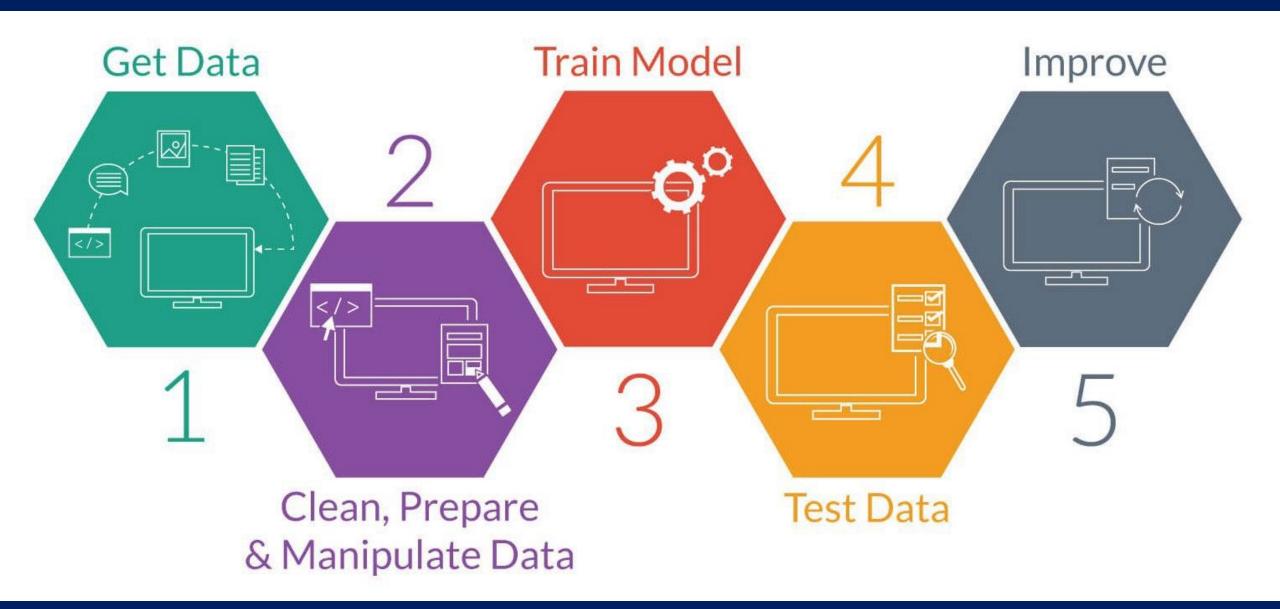
Semi-Supervised Learning

Unsupervised Learning

Look at the label & Output

Machine Learning Algorithms Cheat Sheet





Time to Practice



Reference

- https://www.edureka.co/blog/what-is-big-data/
- https://www.edureka.co/blog/big-data-tutorial
- https://www.edureka.co/blog/big-data-applications-revolutionizing-various-domains/
- https://www.datahen.com/blog/data-scraping-vs-data-crawling/
- https://www.talend.com/resources/what-is-data-preparation/
- https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
- https://towardsdatascience.com/cleaning-and-preparing-data-in-python-494a9d51a878
- https://www.kaggle.com/timoboz/google-trends-data
- https://www.kdnuggets.com/2017/06/which-machine-learning-algorithm.html
- https://medium.com/@pytyagi/work-flow-in-machine-learning-project-327eddb946b4

Image

- https://www.123rf.com/photo_10823919_illustration-of-a-kid-carrying-a-box-full-of-toys-for-donation-or-organizing-ng.html
 ng.html
- https://www.123rf.com/photo_11197789_illustration-of-kids-buying-souvenirs-from-a-toy-store.html?fromid=dGF_50G1ESHhVa0w0UVZodEpvcHJNZz09
- https://www.xenonstack.com/images/blog/10-vs-of-big-data.png
- https://www.uihere.com/free-cliparts/cartoon-clip-art-thinking-woman-1601359
- https://www.uihere.com/free-cliparts/stock-photography-thought-illustration-thinking-child-1388176
- https://www.uihere.com/free-cliparts/big-data-analytics-information-internet-of-things-ppt-element-1330286
- https://www.uihere.com/free-cliparts/school-cartoon-royalty-free-illustration-2-children-holding-up-their-hands-i-n-front-of-the-seat-1329114
- https://www.uihere.com/free-cliparts/idea-drawing-icon-vector-bulb-1267305
- https://id.pinterest.com/pin/541980136404391230
- https://depositphotos.com/stock-photos/dig.html?filter=all&qview=49849527
- https://depositphotos.com/vector-images/detective-cartoon-vector.html?qview=47423759
- http://www.dataception.in/blog/machine-learning/
- https://www.efilecabinet.com/grow-your-business-immediately-with-efilecabinet/im-scared-that-i-have-way-too-much-data-to-migrate-and-even-if-i-could-it-would-take-forever-to-organize/

THANK YOU

Find us on:







