



SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)



Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai
Accredited by NAAC-UGC with 'A++' Grade (Cycle III) & Accredited by NBA (B.E - CSE, EEE, ECE, Mech & B.Tech.IT)
COIMBATORE-641 035, TAMIL NADU

DEPARTMENT OF COMPUTER APPLICATIONS

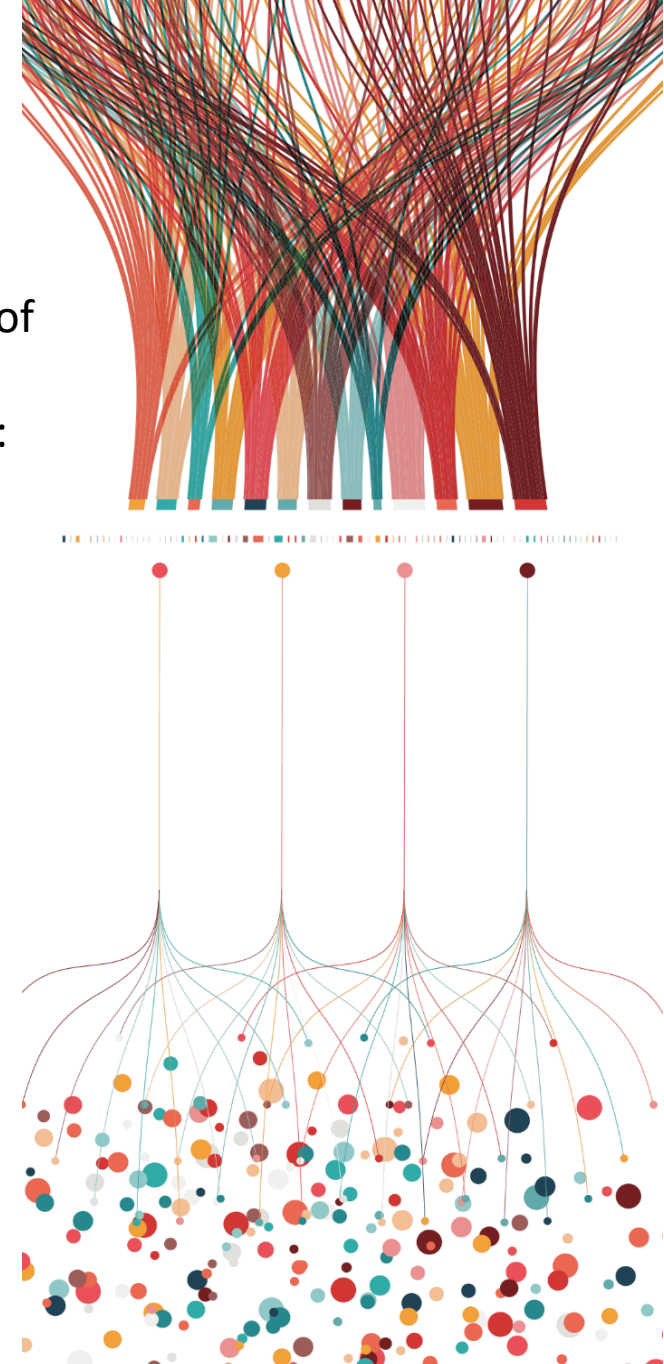
19CAE716 – DATA SCIENCE

UNIT – I: INTRODUCTION TO DATA SCIENCE

TOPIC: FACETS OF DATA

In Data Science and Big Data you'll come across many different types of data, and each of them tends to require ***different tools and techniques***. The main categories of data are these:

- *Structured*
- *Unstructured*
- *Natural Language*
- *Machine-generated*
- *Graph-based*
- *Audio, video and images*
- *Streaming*





MySQL®



Structured Data

- ✓ Structured data refers to data that is organized and formatted in a way that is easily searchable and queryable.
- ✓ It is highly organized and follows a predefined structure, typically in the form of tables with rows and columns.
- ✓ Each piece of data is stored in a specific field within a table, making it easy to retrieve and analyze.

ID	Emp ID	Full Name	Job Title	Department	Job Function	Gender	Manager	Start Date	Salary	Amount	Rate	Country	City	Post Code
1	100017	Emily Davis	Sr. Manager	Research & Development	Female	Blank	35	4/9/2016	\$41,804	10%	United States	Seattle	98106/2021	
2	100420	Theodore Deth	Technical Architect	IT	Manufacturing	Male	Asian	39	11/29/2007	299,975	8%	China	Chongqing	
4	100211	Lynn Sanders	Director	Finance	Specialty Products	Female	Caucasian	35	10/26/2006	\$18,599	20%	United States	Chicago	
5	100282	Penelope Jordan	Computer Systems Manager	IT	Manufacturing	Female	Caucasian	28	9/27/2019	\$68,813	7%	United States	Chicago	
6	100493	Austin Vu	Sr. Analyst	Finance	Manufacturing	Male	Asian	35	11/20/2005	\$56,403	6%	United States	Phoenix	
7	100044	Joshua Gupta	Account Representative	Sales	Corporate	Male	Asian	37	1/28/2012	\$50,981	0%	China	Chongqing	
8	101510	Ruby Barnes	Manager	IT	Corporate	Female	Caucasian	27	7/1/2020	\$19,786	80%	United States	Phoenix	
9	104312	Luke Martin	Analyst	Finance	Manufacturing	Male	Black	25	5/16/2020	\$66,139	6%	United States	Miami	3/20/2021
10	104313	Easton Bailey	Manager	Accounting	Manufacturing	Male	Caucasian	29	1/29/2019	\$13,127	6%	United States	Austin	
11	101018	Maddeline Wyatt	Sr. Analyst	Finance	Specialty Products	Female	Caucasian	34	4/13/2018	\$73,393	6%	United States	Chicago	
12	100091	Savannah Ali	Sr. Manager	Human Resources	Manufacturing	Female	Asian	36	2/12/2009	\$57,333	10%	United States	Miami	
13	103344	Carroll Rogers	Controls Engineer	Engineering	Specialty Products	Female	Caucasian	27	10/21/2011	\$109,851	6%	United States	Seattle	
14	100030	Elis Jones	Manager	Human Resources	Manufacturing	Male	Caucasian	39	2/14/2009	\$101,096	6%	United States	Austin	
15	104229	Everleigh Ng	Sr. Manager	Finance	Research & Development	Female	Asian	31	6/20/2011	\$46,742	30%	China	Shanghai	
16	100166	Robert Wang	Sr. Analyst	Accounting	Specialty Products	Male	Asian	31	11/10/2012	\$92,378	6%	United States	Austin	
17	100049	Isabella Xi	Vice President	Marketing	Research & Development	Female	Asian	41	3/12/2013	\$49,270	30%	United States	Seattle	
18	100163	Bella Powell	Director	Finance	Research & Development	Female	Black	65	2/4/2002	\$73,837	20%	United States	Seattle	
19	100084	Carroll Silva	Sr. Manager	Marketing	Specialty Products	Female	Latino	64	11/7/2003	\$14,838	13%	United States	Seattle	
20	104116	David Barnes	Director	IT	Corporate	Male	Caucasian	64	11/2/2013	\$106,303	24%	United States	Seattle	
21	100605	Adam Tang	Director	Sales	Research & Development	Male	Asian	65	2/9/2002	\$166,331	18%	China	Chongqing	
22	100480	Etsai Alvarado	Sr. Manager	IT	Manufacturing	Male	Latino	36	1/9/2012	\$46,140	30%	Brazil	Miami	
23	104312	Evelyn Lee	Director	Sales	Manufacturing	Female	Latino	36	4/22/2011	\$151,793	21%	United States	Seattle	
24	105484	Logan Rivera	Director	IT	Research & Development	Male	Latino	39	3/24/2002	\$12,787	20%	Brazil	Miami	
25	100671	Lionardo Dixon	Analyst	Sales	Specialty Products	Male	Caucasian	37	3/22/2019	\$65,996	6%	United States	Seattle	
26	100073	Mario Iyer	Vice President	Sales	Specialty Products	Male	Asian	44	12/2/2014	\$87,172	13%	China	Chongqing	
27	103206	Jose Henderson	Director	Human Resources	Specialty Products	Male	Black	41	4/17/2015	\$152,239	23%	United States	Seattle	
28	100045	Abigail Wang	Quality Engineer	Engineering	Specialty Products	Female	Latino	36	2/2/2005	\$96,381	6%	Brazil	Miami	
29	100134	Wynett Chin	Vice President	Engineering	Specialty Products	Male	Asian	43	6/7/2004	\$46,215	10%	United States	Seattle	
30	103463	Carston Lu	Engineering Manager	Engineering	Specialty Products	Male	Asian	64	12/14/2006	\$99,354	12%	China	Chongqing	
31	100304	Dylan Choi	Vice President	IT	Corporate	Male	Asian	63	5/11/2012	\$23,141	34%	China	Chongqing	
32	102094	Esther Kumar	IT Coordinator	IT	Research & Development	Male	Asian	28	6/25/2017	\$84,775	6%	United States	Seattle	

```

nsq.csv_file_sample_2021_04_08.csv - Notepad
File Edit Format View Help
Keyword,Min Monthly Volume,Max Monthly Volume,Specific Monthly Volume,Difficulty,Rank
perfect tower 2 improve chance to find modules,0,10,10.5370846,1,33
straight-forward analytics solution,0,10,1,34
bi-soft,11,50,16.71429414,1,42
vantagens self service,11,50,27.24240562,1,46
turn key analytics,0,10,2,26
southwest virtual agents,0,10,3,34
coming out bi 7 watch online free,11,50,27.24240562,3,36
coming out bi 7 watch online,11,50,27.24240562,3,49
power music login,101,200,121.5019653,4,18
bi rapportages,51,100,58,4,38
legal dashboard,0,10,1,5,18
caso de exito big data netflix,11,50,27.24240562,5,21
consultora de bi,0,10,5,32
power bi consulting london,0,10,6,15
adobe analytics danismanlik,0,10,6,16
spend analytics dashboard,0,10,6,18
syscon online login,11,50,38.68984882,6,23
spend analysis dashboard,11,50,27.24240562,6,24
escalabilidade em saas,0,10,6,27
usa ge november 2019,0,10,10.5370846,6,34

```



Unstructured Data

- ✓ Unstructured data is ***data that isn't easy to fit into a data model*** because the content is context-specific or varying. One example of unstructured data is your ***regular email***.
- ✓ Although email contains structured elements such as the sender, title, and body text, it's a challenge to find the number of people who have written an email complaint about a specific employee because so many ways exist to refer to a person, for example.
- ✓ The thousands of different languages and dialects out there further complicate this.
- ✓ A human-written email, is also a perfect example of natural language data.



Text files and documents



Server, website, and applications logs



Sensor data



Image files



Video files



Audio files



Emails



Social media data



Agent notes



Surveys



Web forms



Mail



Chats

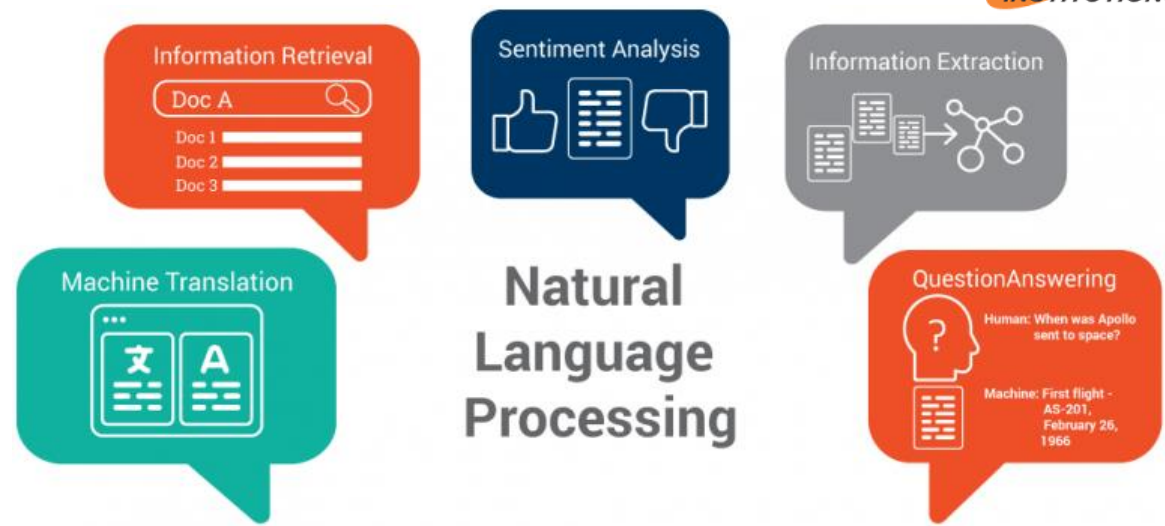


Quality evaluations



Natural Language

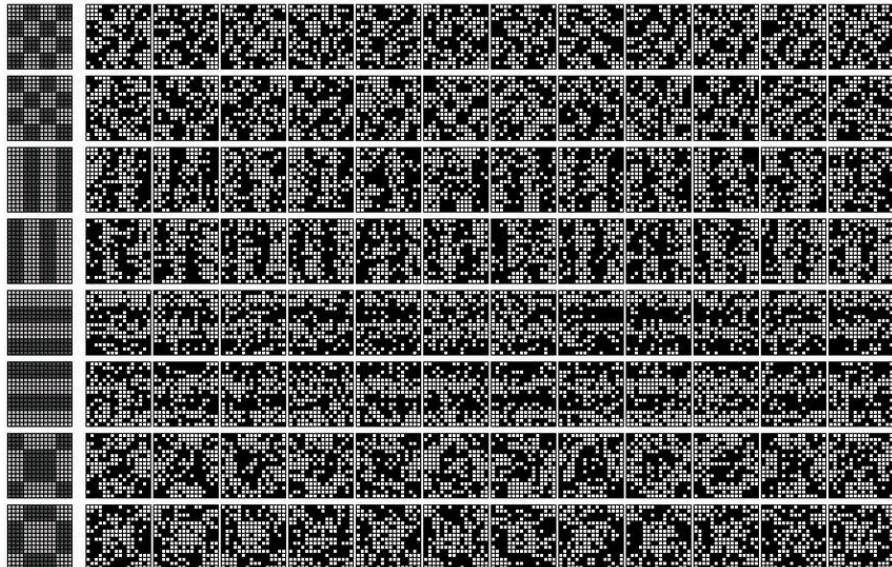
- ✓ Natural language is a **special type of unstructured data**; it's challenging to process because it requires knowledge of specific **data science techniques and linguistics**.
- ✓ The natural language processing community has had success in **entity recognition, topic recognition, summarization, text completion, and sentiment analysis**, but models trained in one domain don't generalize well to other domains.
- ✓ Even state-of-the-art techniques aren't able to decipher the meaning of every piece of text.





Machine – Generated Data

```
CSIPERF:TXCOMMIT:313236
2014-11-28 11:36:13, Info
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info
Beginning NT transaction commit...
2014-11-28 11:36:13, Info
trace:
CSIPERF:TXCOMMIT:273983
2014-11-28 11:36:13, Info
70), objectname [6]"(null)"
2014-11-28 11:36:13, Info
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:13, Info
Beginning NT transaction commit...
2014-11-28 11:36:14, Info
trace:
CSIPERF:TXCOMMIT:386259
2014-11-28 11:36:14, Info
71), objectname [6]"(null)"
2014-11-28 11:36:14, Info
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:14, Info
Beginning NT transaction commit...
2014-11-28 11:36:14, Info
trace:
CSIPERF:TXCOMMIT:375581
```



- ✓ Machine-generated data is informative that's ***automatically created by a computer, process, application or other machine without human intervention.***
- ✓ Machine-generated data is becoming a major data resource and will continue to do so.
- ✓ ***The analysis of Machine data relies on highly scalable tools, due to high volume and speed.***



Audio, Images, & Videos

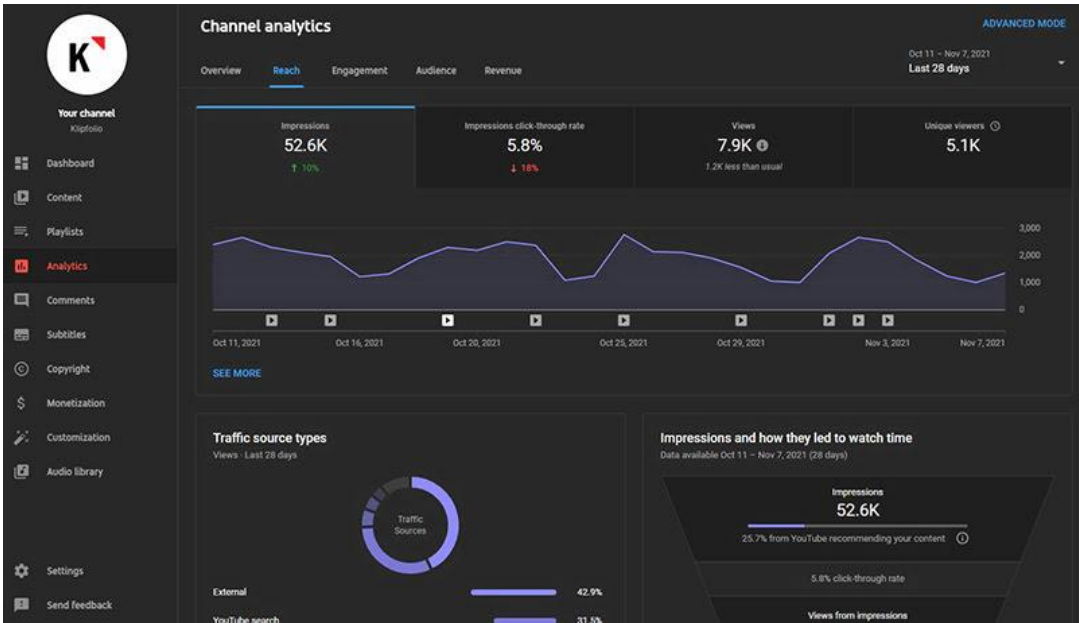


- ✓ Audio, image, and video are data types that pose specific challenges to a data scientist.
- ✓ Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.
- ✓ Multimedia data in the form of audio, video, images and sensor signals have become an integral part of everyday life.
- ✓ Moreover, they have revolutionized product testing and evidence collection by providing multiple sources of data for quantitative and systematic assessment.





Streaming Data



- ✓ While streaming data can take almost any of the previous forms, it has an extra property.
- ✓ The ***data flows into the system when an event happens instead of being loaded*** into a data store in a batch.
- ✓ Although it isn't really a different type of data, we treat it here as much because you need to adapt your process to deal with this type of information.
- ✓ Examples are the “What’s trending” on Twitter, live sporting or music events and the stock market.