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

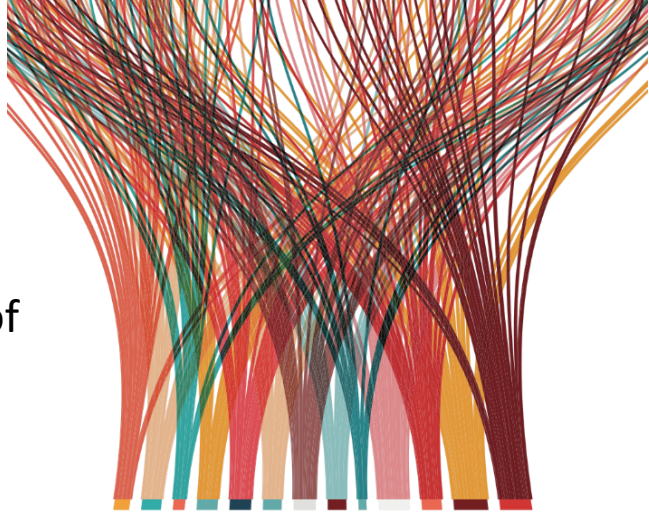
19CAE716 – DATA SCIENCE

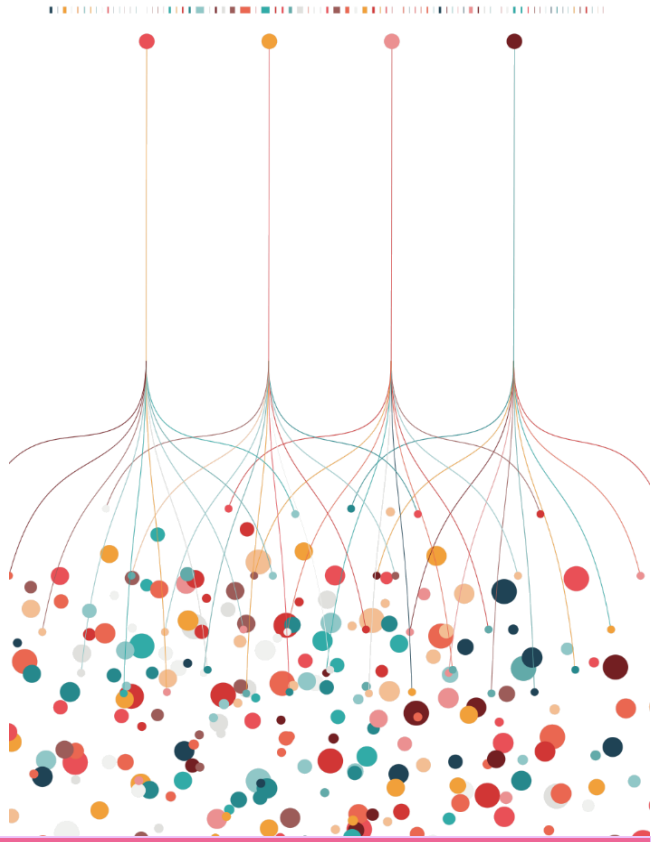
UNIT – I: INTRODUCTION TO DATA SCIENCE

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:

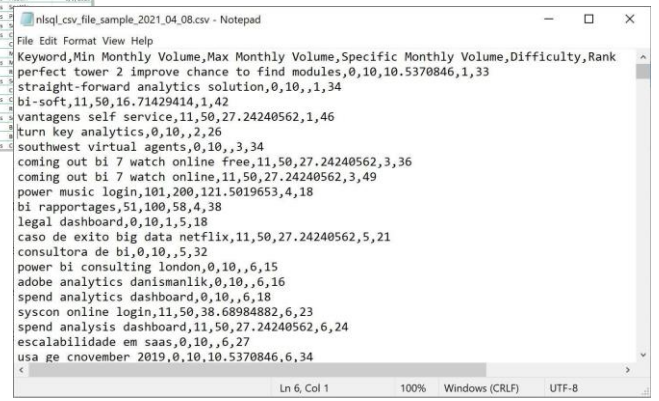
- *Structured*
- *Unstructured*
- *Natural Language*
- *Machine-generated*
- *Graph-based*
- *Audio, video and images*
- *Streaming*



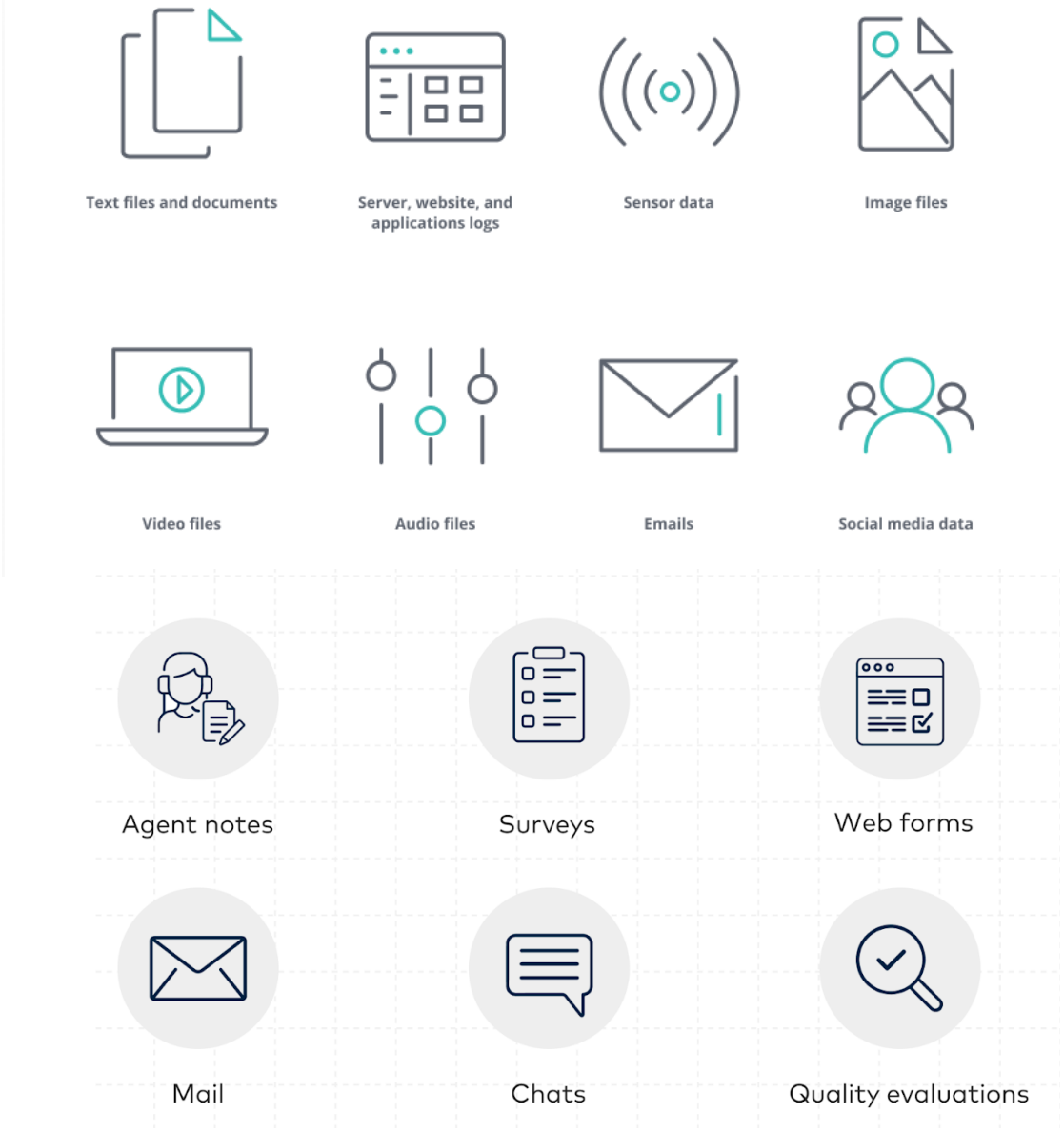


Structured Data

- ✓ 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.



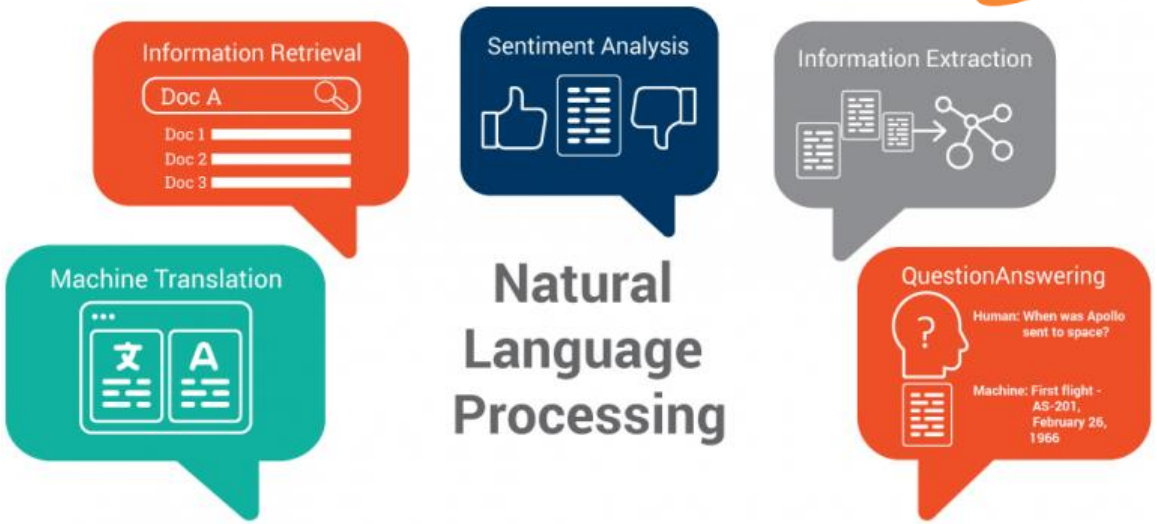
Unstructured Data



- ✓ 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.

Natural Language

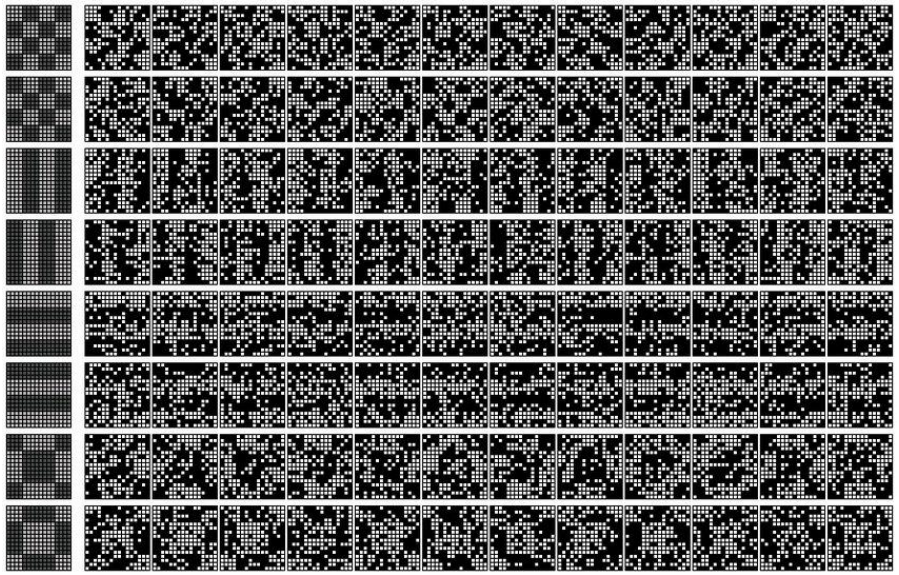
- ✓ Natural language is a ***special type of unstructured data***; it's challenging to process because it requires knowledge of specific ***data science techniques and linguistics.***
- ✓ The natural language processing community has had success in ***entity recognition, topic recognition, summarization, 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.



alexa

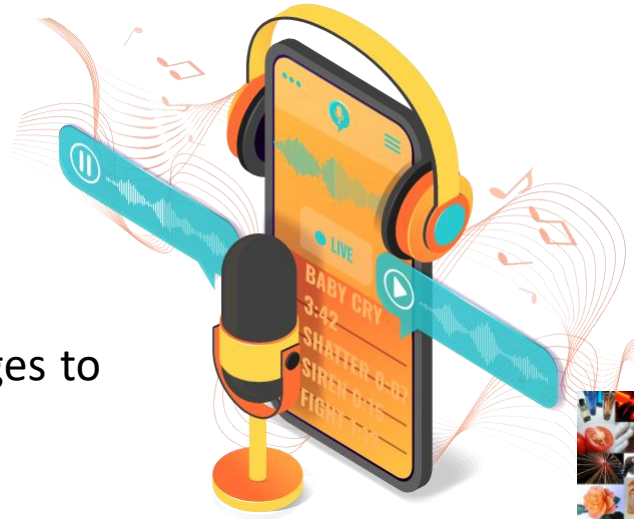
Machine – Generated Data

```
CSIPERF:TXCOMMIT;313236
2014-11-28 11:36:13, Info                  CSI    00000153 Creating NT transaction (seq
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000154 Created NT transaction (seq 69)
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info                  CSI    00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...
2014-11-28 11:36:13, Info                  CSI    00000156@2014/11/28:10:36:13.705 CSI perf
trace:
CSIPERF:TXCOMMIT;273983
2014-11-28 11:36:13, Info                  CSI    00000157 Creating NT transaction (seq
70), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000158 Created NT transaction (seq 70)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:13, Info                  CSI    00000159@2014/11/28:10:36:13.764
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015a@2014/11/28:10:36:14.094 CSI perf
trace:
CSIPERF:TXCOMMIT;386259
2014-11-28 11:36:14, Info                  CSI    0000015b Creating NT transaction (seq
71), objectname [6]"(null)"
2014-11-28 11:36:14, Info                  CSI    0000015c Created NT transaction (seq 71)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:14, Info                  CSI    0000015d@2014/11/28:10:36:14.106
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015e@2014/11/28:10:36:14.428 CSI perf
trace:
CSIPERF:TXCOMMIT;375581
```



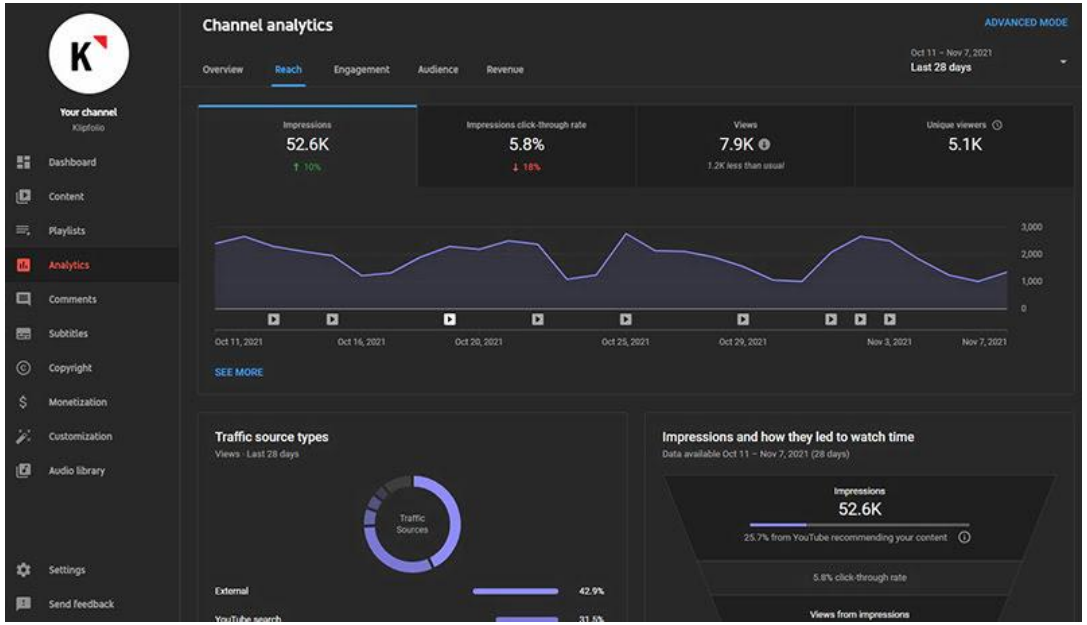
- ✓ Machine-generated data is informative that's ***automatically created by a computer, process, application or other machine without human intervention.***
- ✓ Machine-generated data is becoming a major data resource and will continue to do so.
- ✓ ***The analysis of Machine data relies on highly scalable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