# MLT-06 Summary Document

## Overview

This document summarizes MLT-06 lecture on Natural Language Processing and Sentiment Analysis in Trading. This lecture talks about alternative data, tools for processing it, various processing techniques and packages to implement it.

The lecture covers the following topics:

- Alternative data
    - What is alternative data?
    - Origin of demand for alternative data
    - Categories of data
    - Spending pattern
    - Usage
- Content and tools
- Text processing
    - What is text processing?
    - Tools for text processing
    - List of activities in text processing
- Topic modelling
    - Latent Semantic Analysis (LSA)
    - Latent Dirichlet Allocation (LDA)
- Machine readable news
- Commercial packages

## Alternative Data

### What is alternative data?

Alternative Data refers to the data that are beyond “market discovered” trades and quotes data. The data with high diversity, variety, size, frequency and volume is considered Big Data.

## Origin of demand for alternative data

- Hedge Funds + Quant Funds + Algo Trading Houses - Use Alt Data to obtain hidden state information not available via traditional sources
- Corporates, Venture Capitalists use Alt data to get an edge on their decision making

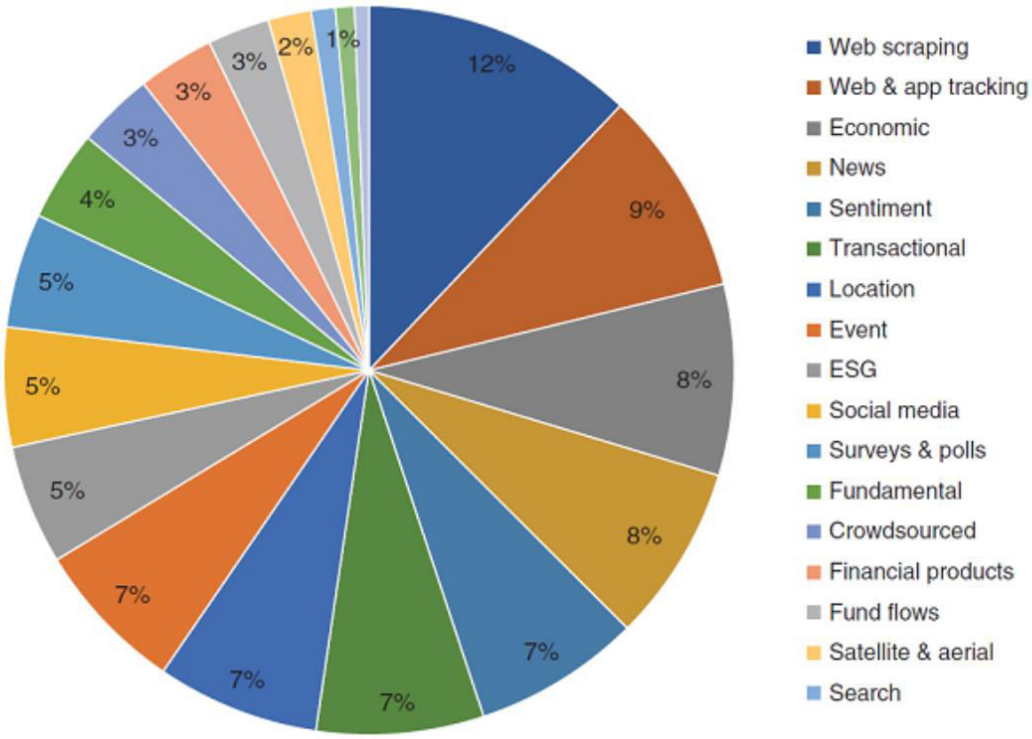
## Categories

*The following snapshot shows the multiple categories of alternative datasources*


*Source:* Neudata

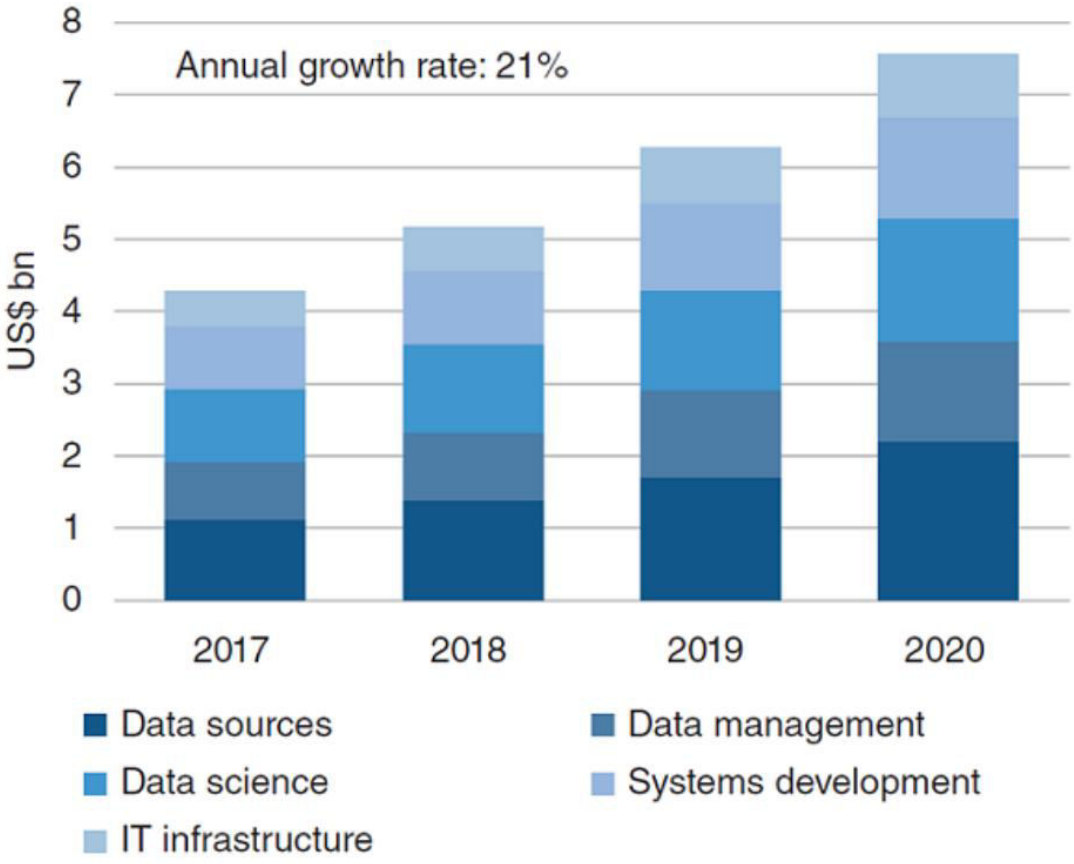
## Usage patterns

*The following snapshot shows the distribution of usage pattern of various alternative data sources*




## Spending trends

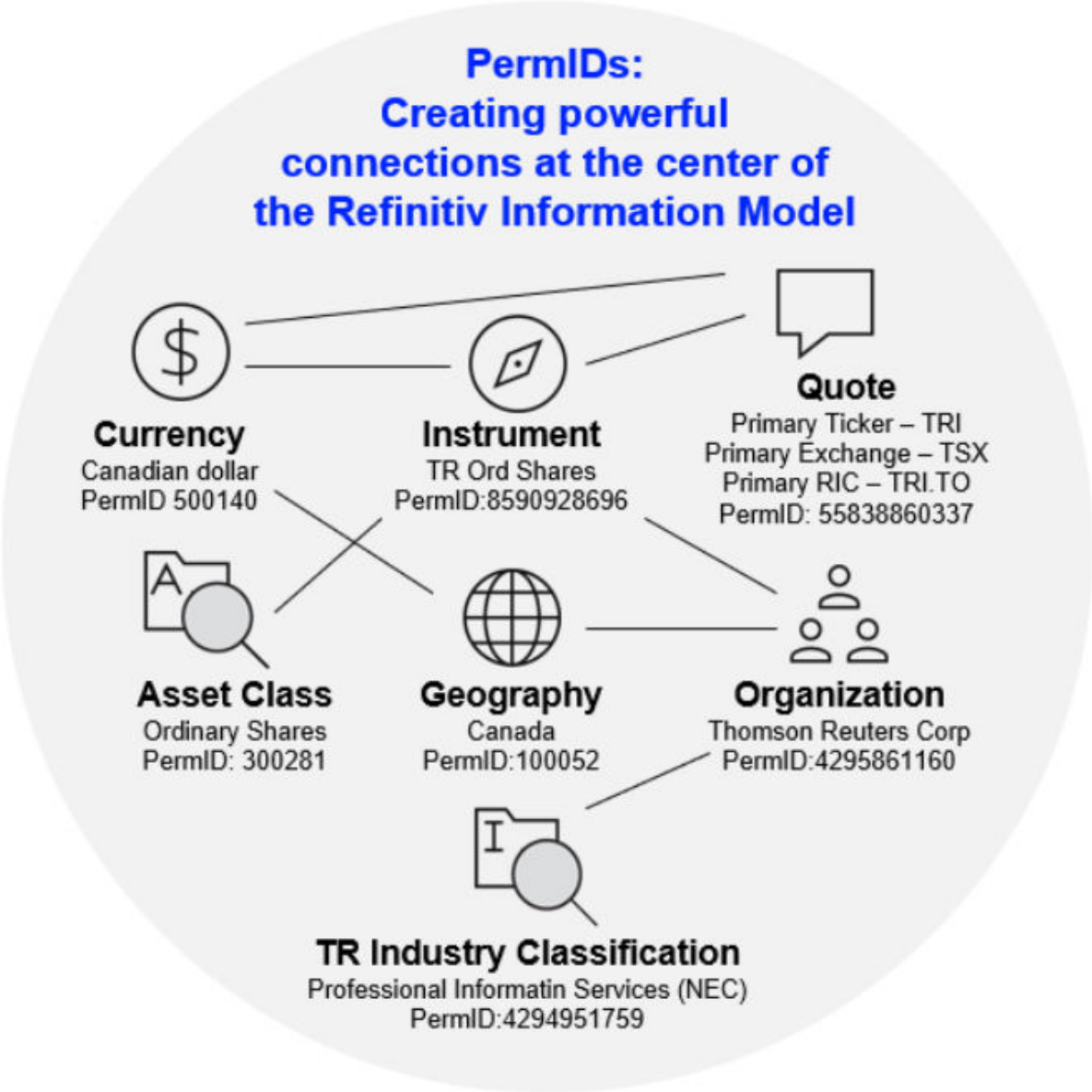
The following snapshot shows the spending trends of alternative datasets in various fields.




**Content and tools**

- Permid.org is an open data content tool, which gives an id to each entity, known as PERMid.
- The entity could be an organization, instrument, orquote, etc. All the related information (directors, addresses, stakeholders etc.) is linked with these entities.
- The PERMid is an identifier that serves as a single reference for all the unstructured data that you come across.
- PermID essentially maps external unstructured data to structured data.

*The following snapshot shows an example of PermID mapping*




**Use cases of PermIDs**

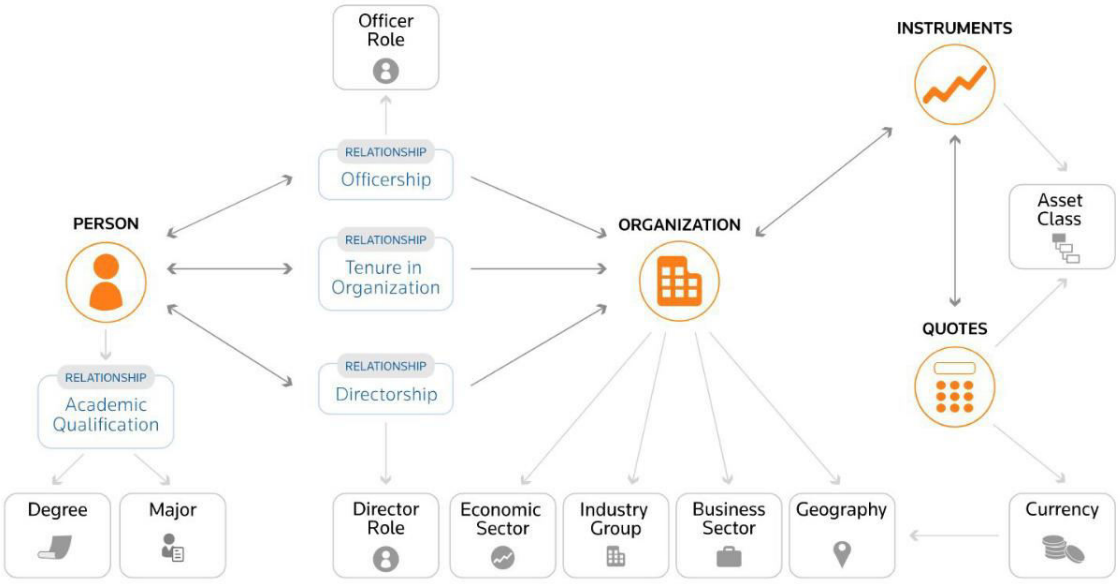
- Developing platforms to easily search and connect your data, uncovering the right connections.
- Tagging all of your data and exposing powerful linkages for unique insights.

- Co-mingle internal and external data to manage risks

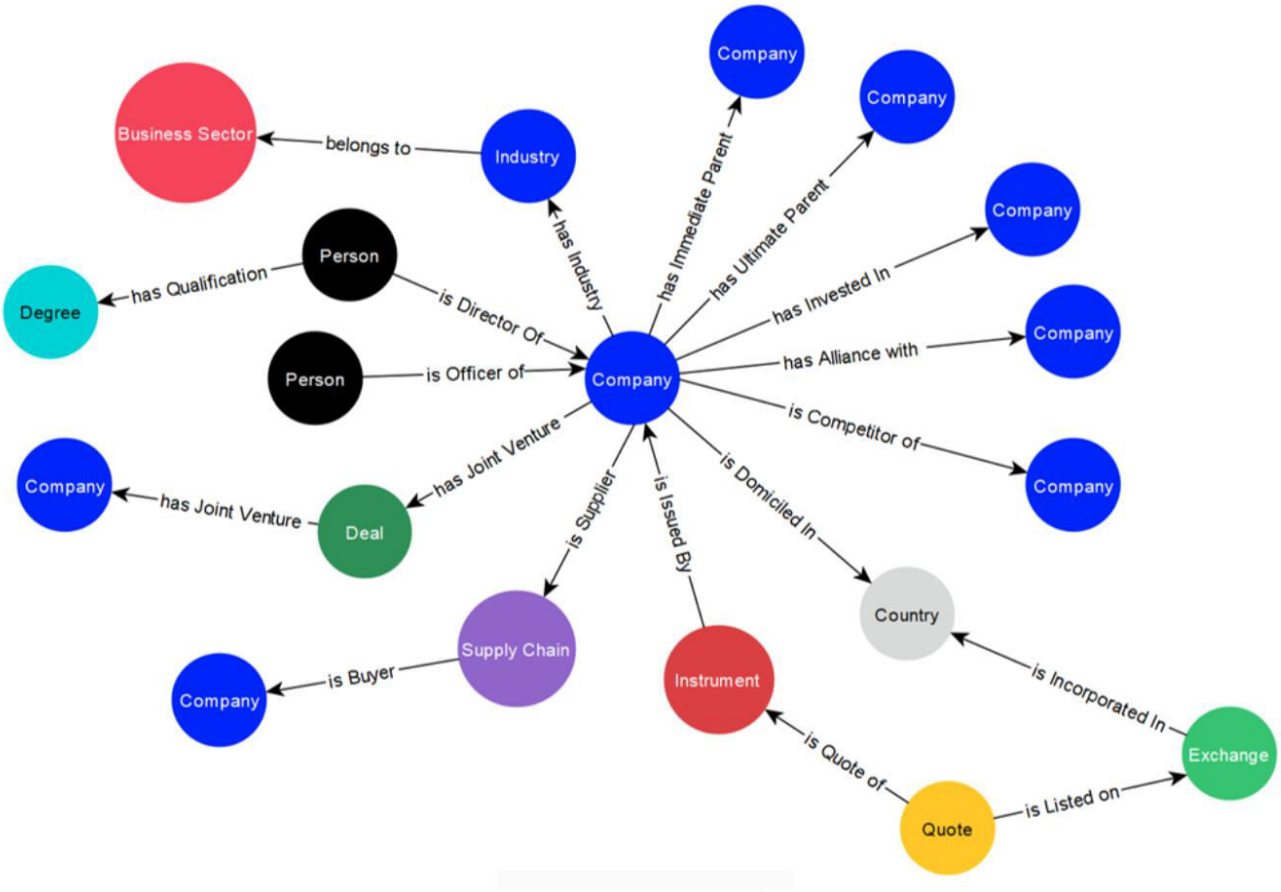
**Knowledge graph**

Knowledge graphs are tools to represent data in a graph-structure.

*The following snapshot shows a knowledge graph to explain the organization of data*



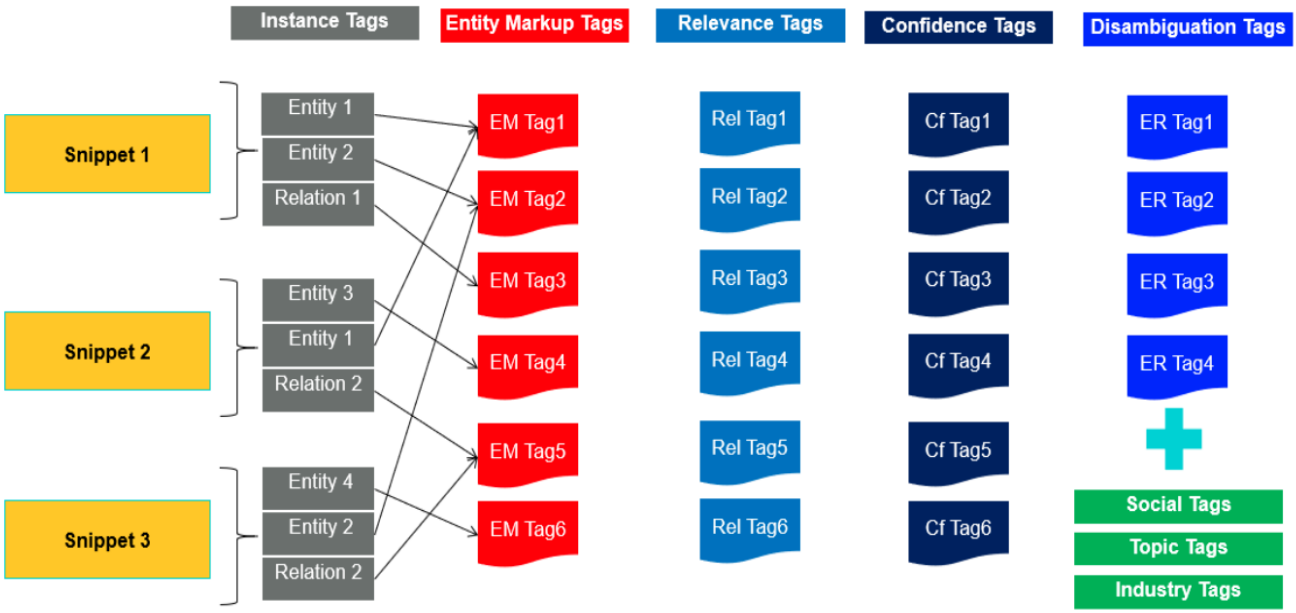

*The following snapshot shows any example of a knowledge graph for a company*




**Intelligent Tagging**

- It turns qualitative, unstructured text into quantitative and actionable insight.
- Inputs can be text/html, text/xml, application/pdf, text/raw
- It supports English, French and Spanish

*The following snapshot shows the working of the intelligent tagging tool.*




- Entities can be tagged by the engine like city, company, country, etc.
- Relations can be tagged by the engine such as Buybacks, Acquisition, etc.

## Text Processing

### What is text processing?

Text Processing is an important stage in algorithm modelling that uses natural language processing (NLP) to transform the unstructured text into normalized, structured data suitable for ML algorithms to understand and process.

### Tools for text processing

spaCy is one of the popular tools in the NLP domain. It is an open-source library for advanced Natural Language Processing (NLP) in Python. And it is free to use.

List of activities in text processing include -

| **Activity** | **Description** |
|---|---|
| Case conversion | Change all the text to a particular case |
| Punctuation | Removing punctuation symbols |
| Numbers | Removing numbers |
| Stopword removal | Remove words like on, to, etc. |
| Stemming | Reduce words to its stem. |
| Tokenization | Segmenting text into words, punctuation marks etc. |
| Part-of-speech (POS) Tagging | Assigning word types to tokens, like verbs or nouns. |
| Dependency Parsing | Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object. |
| Lemmatization | Assigning the base forms of words. For example, the lemma of “was” is “be”, and the lemma of “rats” is “rat”. |
| Sentence Boundary Detection (SBD) | Finding and segmenting individual |

| | |
|---|---|
| | sentences. |
| Named Entity Recognition (NER) | Labelling named “real-world” objects, like persons, companies or locations. |
| Entity Linking (EL) | Disambiguating textual entities to unique identifiers in a knowledge base. |
| Similarity | Comparing words, text spans and documents and how similar they are to each other. |
| Text Classification | Assigning categories or labels to a whole document or parts of a document. |
| Rule-based Matching | Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions. |
| Training | Updating and improving a statistical model’s predictions. |
| Serialization | Saving objects to files or byte strings. |

Some of the recent useful NLP pre-trained models are -

- Universal Language Model Fine-tuning (ULMFiT)
- Embedding from Language Models (ELMo)
- Transformer
- Universal Sentence Encoder (USE)
- GPT-3
- Bidirectional Encoder Representations from Transformers (BERT)
- Transformer-XL
- XLNet

**Topic Modelling**

In the NLP hierarchy of elements, the hierarchy for a document from bottom to top is words-->paragraph-->document->topic. The topic is the most effective element to understand the context from the collection of texts.

The process of recognizing, identifying and extracting the topics from the collection of documents is called topic modelling. LSA and LDA are the two popular techniques in topic modelling.

**LSA (Latent Semantic Analysis)**
It identifies the patterns and finds out relevant and important information from the text document. It uses an unsupervised approach.

It is a very helpful technique in the reduction of dimensions of the matrix or topic modelling by grouping together all the words that have a similar meaning.

**LDA (Latent Dirichlet Allocation)**
It forms clusters or group sentences having the same contextual meaning from the input documents or sentences.

## Machine-readable news

News and sentiment analysis is an important part of Alternate data analysis.

There are some aspects of the news that needs to be considered for trading:

- **News alerts** are as important as news articles. One sentence or phrase of news cannot be ignored if it is relevant.
- **Timestamp -** Sequence of news is very important, particularly it needs to be compared from the same source if it is a small interval. The first alert is also important to track the article and subsequent alerts.
- **Topic Codes -** Describe the news item's subject matter. These cover asset classes, geographies, events, industries/sectors, and other types.
- **Product Codes -** Identify which desktop news product(s) the news item belongs to.
- **Named Item Codes** - Identify the news items that follow a pattern, also called recurring report codes.

- **Attributions -** Organizations that published the news item.

**Commercial Packages**

- Real-Time-News: Reuters News
- Regions: AMERS, EMEA, APAC, Global, US, Europe, Canada, APAC
- News Type: Political / General / Economic News
- Delivery Options: Real-Time Delivery and Historical News Archive