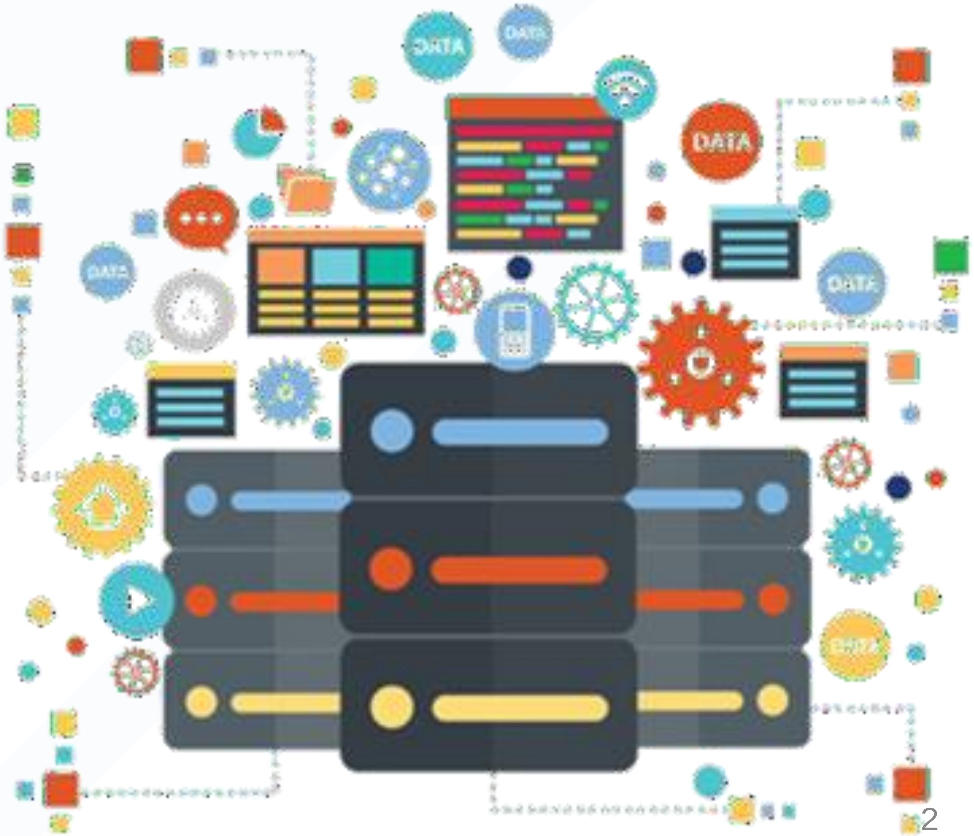
NYP | School of Information Technology

Extract Text Data

Topic Modelling & Sentiment Analysis

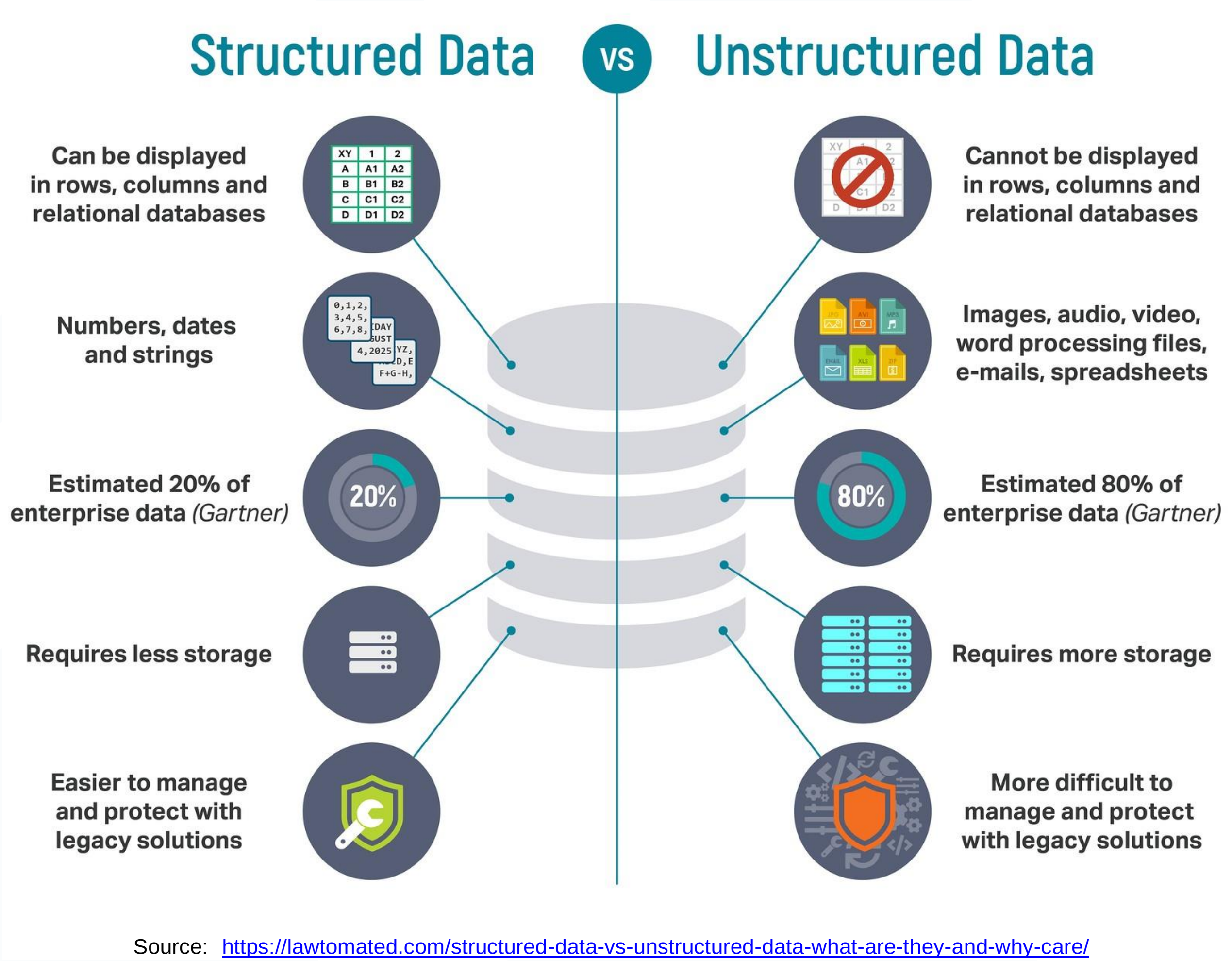
What we will cover

- Data Categories
- Data Sources
- Data Types



Data Categories

Bloom in Unstructured Data


Structured Data
VS
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

Source: https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

About 80% of big data is unstructured data - text, speech, image and video.

How can we extract value from this massive and high growth asset?

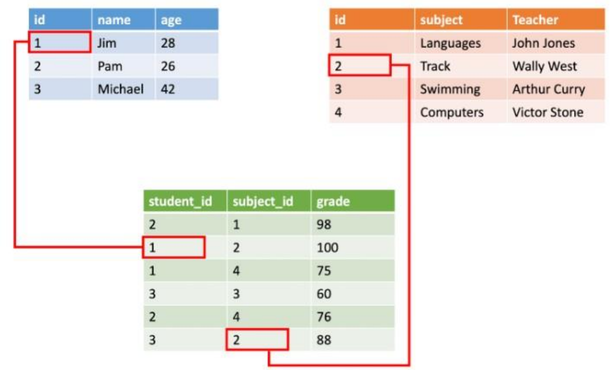
Pros and Cons

Structured

- Pros
  - Easily used by ML algorithms
  - Easily used by business users
  - Accessible by more tools
- Cons
  - Limited usage
  - Limited storage options

Unstructured

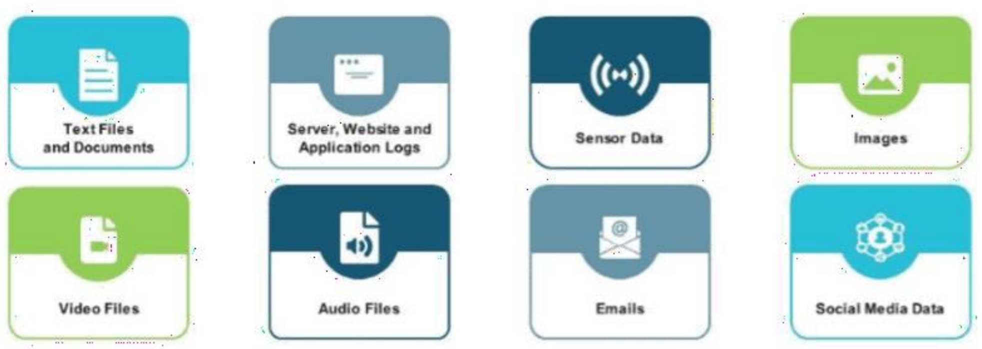
- Pros
  - Native format
  - Fast accumulation rates
  - Data lake storage
- Cons
  - Requires expertise
  - Specialised tools



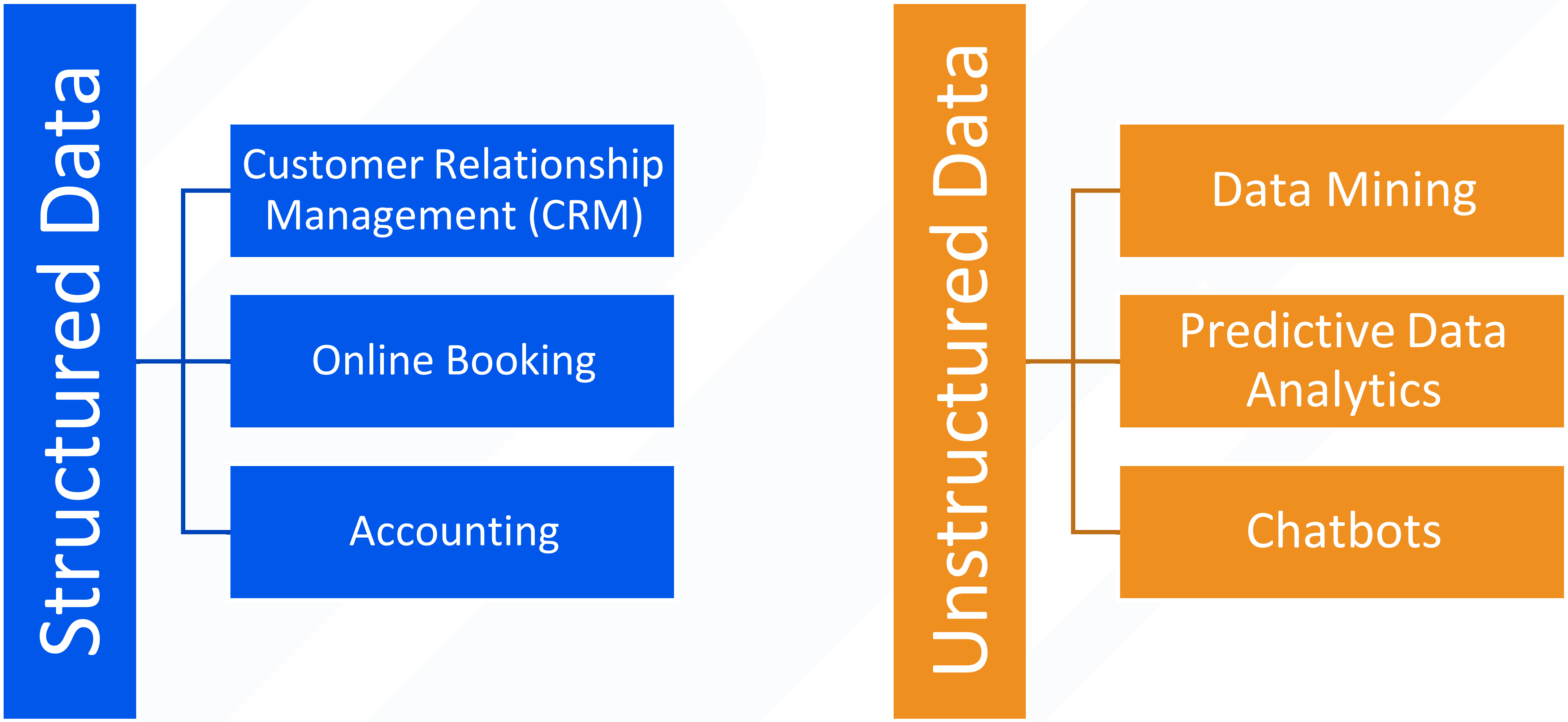
id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88



Uses Cases



Structured Data

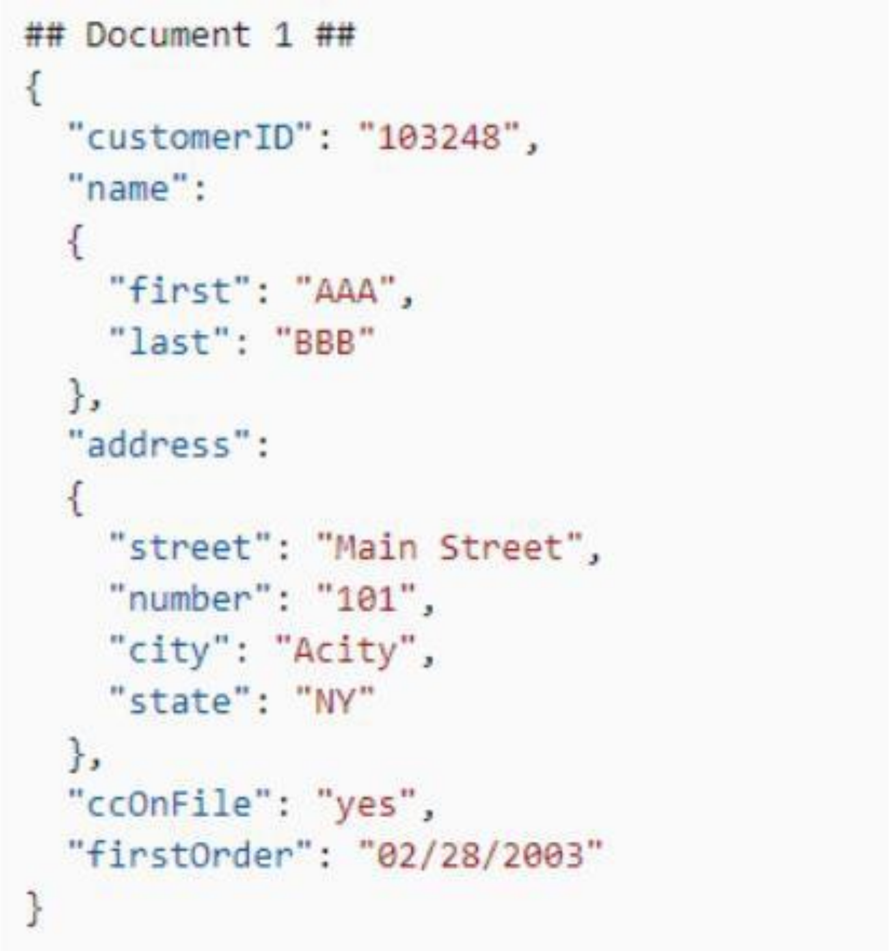
- Customer Relationship Management (CRM)
- Online Booking
- Accounting

Unstructured Data

- Data Mining
- Predictive Data Analytics
- Chatbots

Semi-Structured Data

- “Bridge” between structured and unstructured data
- Does not have predefined data model
- More complex than structured data
- Easier to store than unstructured data
- Uses “metadata” to identify data characteristics and scale data into records and preset fields



```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Source: Structured vs Unstructured Data | Semi Structured Data

Data Sources

Client Data

- Organisation’s own data
- Data storage
  - SQL databases
  - Hadoop clusters
  - Cloud storage
  - Flat files

Free Source

Freely available over the internet

- Free APIs like Twitter
- Wikipedia
- Government data (e.g. http://data.gov.sg)
- Census data (e.g. http://www.census.gov/data.html)
- Health care claim data (https://www.healthdata.gov/)

Web Scraping

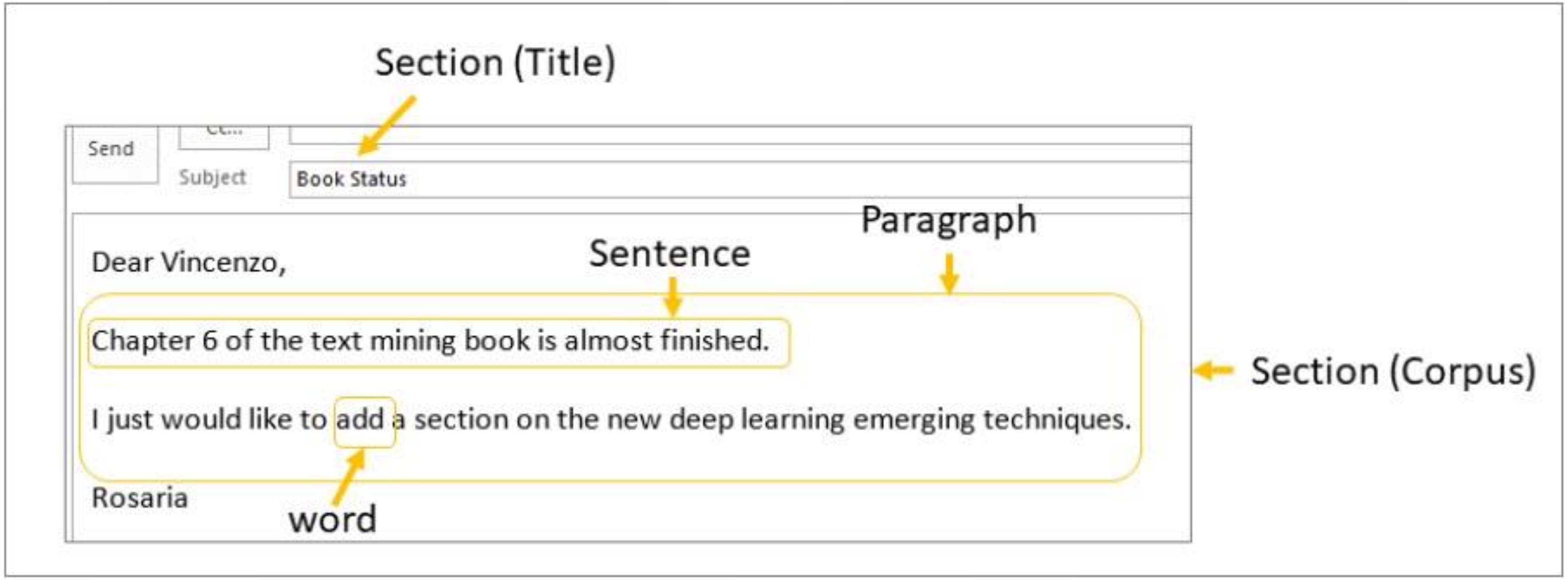
- Extract content/data from
  - Websites
  - Blogs
  - Forums
  - Retail websites for reviews

WEB SCRAPING
HTML WEBSITES
WEB SCRAPING
DATA

Data Types

Data Types for Text Processing

- **Document**: metadata such as title, author(s), source and class are added to original text
- Hierarchical structure of text items: sections, paragraphs, sentences, and words ➔ also referred to as tokens



Subject: Book Status

Dear Vincenzo,

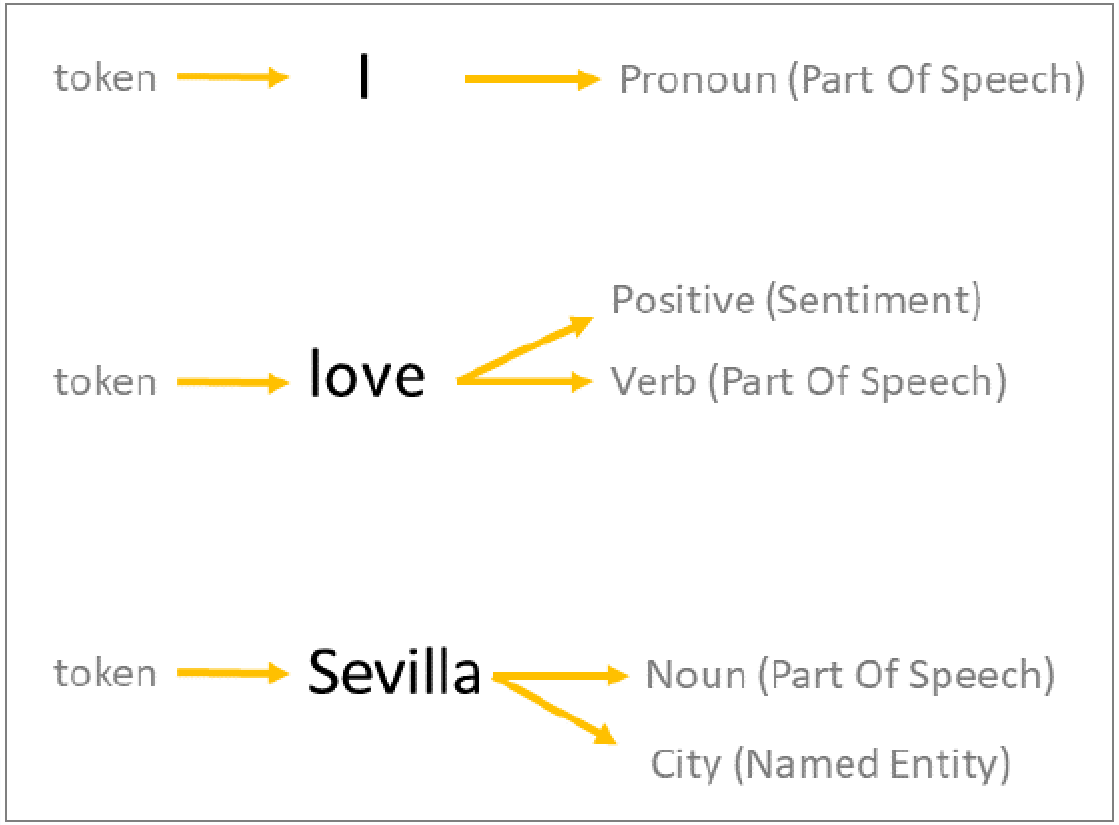
Chapter 6 of the text mining book is almost finished.

I just would like to add a section on the new deep learning emerging techniques.

Rosaria

Data Types for Text Processing

- **Term**: a token with addition of related metadata, and tags
- Tags describe sentiment, part of speech, name entity (if any)



Text Analytics Glossary

Documents and Tokens

Term	Definition
Token	A string of characters representing a unit of text data, also known as a unigram
Bigram	Two tokens in succession E.g.: “New York”
Trigram	Three tokens in succession E.g.: “New York City”
N-gram	N tokens in succession
Corpus	A collection of documents
Corpra	Plural form of corpus

Text Corpora

- Large structured collection of texts or textual data
- Primary purpose ➔ linguistic and statistical analysis
- Annotated with rich metadata

- Some of the Popular Corpora:
  - *Brown Corpus*: million-word corpus for English Language
  - *WordNet*: semantic-oriented lexical database for English Language
  - *Penn Treebank*: consists of tagged and parsed English sentences including annotations like POS tags and grammar-based parse tree
  - *Google N-gram Corpus*: n-gram files up to 5-grams for each language
  - *Reuters Corpus*: collection of Reuters news articles and stories

Any Questions?

We have covered:

- Data Categories
  - Structured vs Unstructured
  - Semi-Structured
- Data Sources
  - Client Data
  - Free Source
  - Web Scrapping
- Data Types
  - Document
  - Term