DATA SCIENCE HOLODECK

Product Development Methodology

by Todorka Dimitrova, PhD

tdi@cphbusiness.dk

cphbusiness
COPENHAGEN BUSINESS ACADEMY

INDEX

ABOUT

The R&D project Data Science Holodeck developed at Copenhagen Business Academy in the period 2021-2023 investigates and integrates the recent advances in AI and VR for accommodating them into the education and business practice of operating with large and complex data and abstract concepts and converting them into valuable knowledge.

The workflow, presented here, describes the methods, models, tools, and development environments, that can be used for creating software applications that automate data processing, analysis, and visualisation, by means of AI and VR.

> *"A workflow consists of an orchestrated and repeatable pattern of activity, enabled by the systematic organization of resources into processes that transform materials, provide services, or process information."* Wikipedia

The workflow is designed, validated, and implemented as a pipeline of operations, that can be configured and orchestrated to create a runnable application of the Holodeck concepts. It shines, applied to user scenarios that involve

- too much data and information (TLDR concept)
- too many sources, multiple dimensions, or huge variety of input formats
- high complexity or abstraction of processing logic and analytical algorithms.

A demo application, illustrating the workflow is presented separately.

The project team consists of Todorka Dimitrova and Jon Bertelsen, lecturers at Copenhagen Business Academy.

TASK FRAME

Objectives

While processing big and complex data or abstract concepts, built of multiple composable pieces of knowledge, either automatically or with human in the loop, people can get overloaded with information and relations, distracted, slow, tired, or inaccurate. It may cause them making errors or taking wrong decisions.

Figure 1. Work with data and documents can be stressful, Source

The problem increases with the variety of data sources, communication channels, and file formats. Integrating diverse data sets can be challenging. Inconsistent data structures and data silos can hinder the effective processing the data.

SMEs, for example, can face different challenges in data processing, despite having smaller datasets compared to larger enterprises. Some of the common problems they may encounter include lack of expertise or resources to ensure proper choice of analytical instruments, quick and quality data processing and insights interpretation.

Learners at education face different problems of similar nature. They accommodate knowledge and skills that require attention, analytical thinking, good memory, and creativity. People learn differently, but all involve their senses and communication abilities in support of the cognition – visual, aural, verbal, physical, social. In traditional 2D data analytics many natural resources of learning are left passive and isolated.

Figure 2. People learn differently, Source

The AI-VR methodology addresses such challenges. It proposes user-friendly analytical instruments, which could facilitate the natural human abilities of sensing and comprehension by revealing, structuring, and interacting with the knowledge in a specific domain (area of interest).

These instruments will not replace the investment in learning and training, but will enhance individuals' data literacy, understanding, and processing capabilities, and will prepare them for making choices that are relevant to their real needs, saving time and waste of resources.

Tasks Definition

The process starts with definition of the domain context and the learning goals – both are human-driven activities. As they are essential for the success of the whole process, humans can be inspired and advised by experimenting with our primers and demo cases.

Once they get clear idea of what they would try to achieve by AI and VR technologies, they are able to follow the technical workflow.

Seen in a nutshell, the process of achieving the objectives looks like the image on Figure 3.

Technically, the workflow decomposes into three stages, each including various types of activities:

1. Ingestion: Domain specific source data is collected and submitted as an input to the processing modules
2. Processing: Data processing operations run selected AI algorithms for analysis and transformation of the input data, which also generate and output insights out of it.
3. Visualisation: The insights are visualised in interactive 3D and VR space.

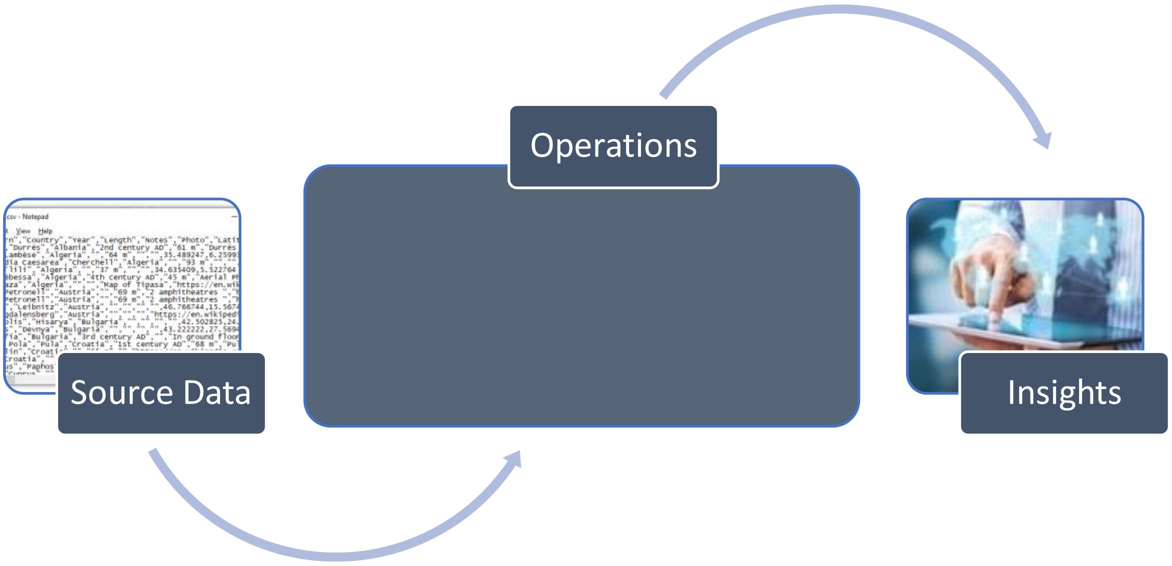


Figure 3. Workflow with a black box of operations, Source author

WORKFLOW STAGES

The solution involves building and applying proper operations on the available data that can generate and visualise useful insights.

We divide and run the operations in five sequential stages, as seen on Figure 4.

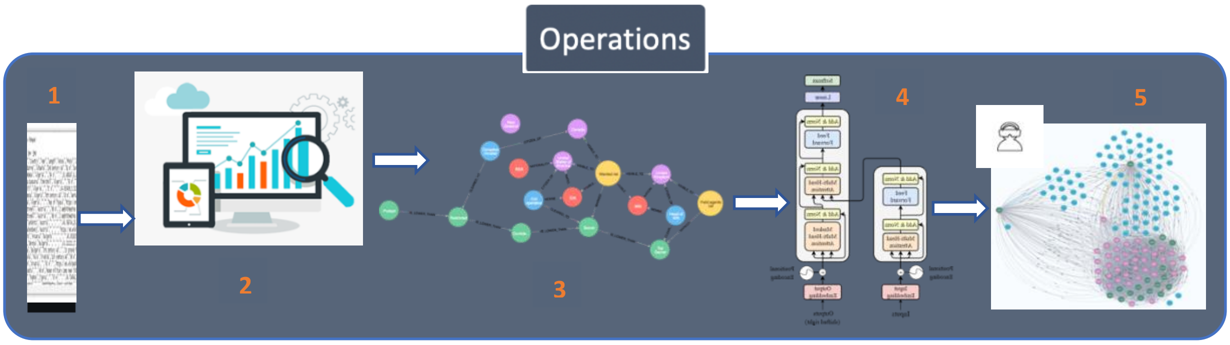


Figure 4. Operations workflow, Source author

Stage 1: Data ingestion
Stage 2: Pre-processing
Stage 3: Building knowledge graph
Stage 4: Processing
Stage 5: 3D/VR Visualisation

Some of the stages have alternative implementations, others produce outcomes and can interupt the workflow.

Stage 1: Data Ingestion

1.1 Data Acquisition

The procedure starts with collecting the sources of data and documents that may contain information relevant to building a picture of the domain.

These can be document files in different format, e-mail, web pages, wiki data, public and private APIs, video clips and audio podcasts.

Figure 5. Data sources

Each type of documents mentioned above has a connector and a loader, that read the data stored in the file and transfer it into a data structure.

Data structures are aggregated as appropriate and stored for further processing.

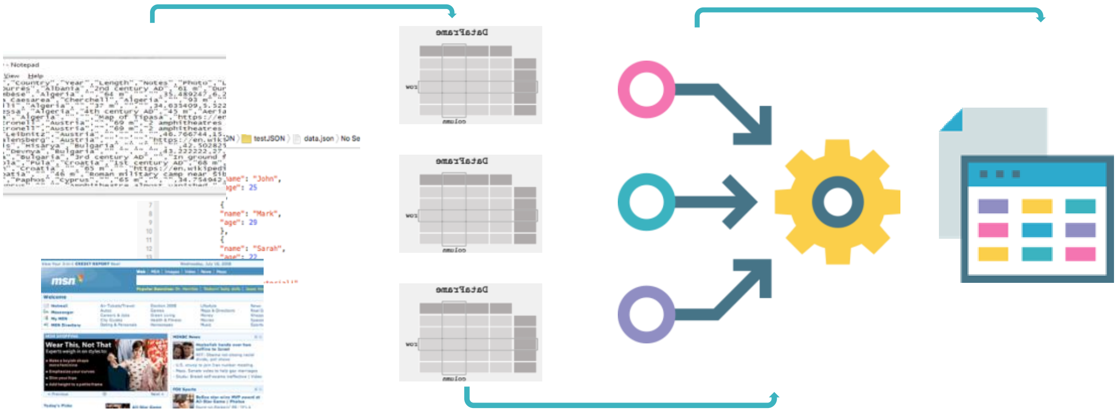


Figure 6. Data ingestion Source author

The project Data Science Holodeck uses both a home created Python library of readers and tools from LangChain framework.

1.2 Data Cleaning and Exploration

At this step, we explore the collected data to get oriented in the available quantity and quality.

Ususly, much of data is noisy and garbage – meaning not ready for automation of operations. GIGO (Garbidge In, Garbidge Out) is a slang in AI/ML, used to denote the importance of the quality of available data. Cleaning techniques are applied as much as they are needed.

Methods from descriptive statistic and diagramming visualisation are applied to enable clearer observation and accurate estimation of the characteristics of data samples, in order to decide on further processing strategies.

Based on the exploration, we decide how to persist the available data, what kind of data is still missing and where it can be found, how to transform the structures, what king of preprocessing is needed.

Figure 7. Data Exploration

The project Data Science Holodeck uses the standard Python packages and libraries for structuring, cleaning, and exploration of data.

Stage 2: Data Preprocessing

A large part of the workflow's duration and operations go to the pre-processing of the source data to an acceptable for AI analysis format.

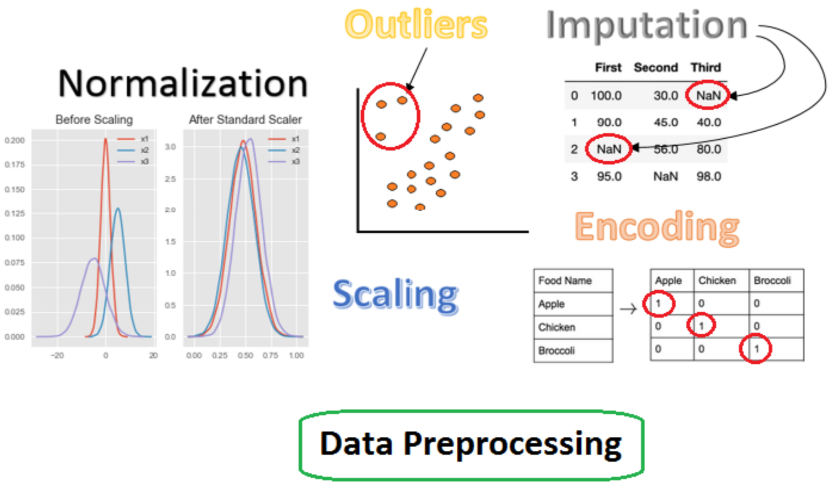
Different types of engineering and conversion applies to the different data types, in dependence with the goals of the analysis and properties of the available resources.

2.1 Numeric Data

For example, the numeric and categorical data is

- examined and processed for completion (filling in missing values with values calculated after a specific rule)
- cleaned by noise, e.g. removed outliers
- transformed and reengineered – normalised, deduplicated, and reduced in size (number of attributes)

Methods of inferential statistics are applied for testing hypotheses and drawing conclusions about the whole population based on testing the available data sets.



	First	Second	Third
0	100.0	30.0	NaN
1	90.0	45.0	40.0
2	NaN	56.0	80.0
3	95.0	NaN	98.0

Food Name
Apple
Chicken
Broccoli

→

Apple	Chicken	Broccoli
1	0	0
0	1	0
0	0	1

Figure 8. Data Pre-processing, Source

PCA (Principal Component Analysis)

An important step in data engineering is so called PCA (Principal Component Analysis) – a dimensionality-reduction analytical procedure, where a large set of data variables (attributes) is transformed into a smaller set in a way, which preserves the original quantity of information sufficiently.

It means, with trading little accuracy, PCA gains much more simplicity and fast performance of ML algorithms.

Preparing the data for analysis is an iterative process. It can continue repeating certain steps until the data and the selected models perform test analysis with sufficient accuracy and precision of results.

The project Data Science Holodeck uses ML packages and libraries offered by Python.

2.2 Natural Language

The natural language text data is put under different categories of pre-processing operations, such as:

Language Detection

The original language of input documents determines the procedures and quality of operations to high extend. Being different in vocabulary, the language-specific operations may follow different lexical and semantic rules, abbreviation interpretation, and punctuation elimination. The languages also consider differently the influence of different stop-words (commonly used words, filtered out as insignificant for the meaning of the sentences).

Anonymisation

As soon as the language is recognized, the text is scanned for allocation sensitive words, normally identification and names of people, locations, dates, codes, and similar personal and protected data.

Tokenization

In NLP, tokenisation is a process of splitting a stream of textual data into smaller units: words, terms, sentences, symbols, or some other meaningful elements called tokens.

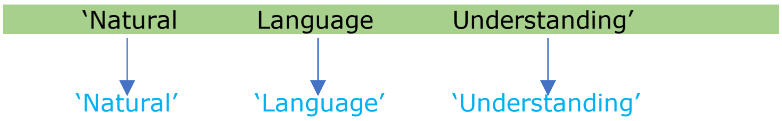


Figure 9. Splitting into word tokens

The process of segmentation is automated and uses the white spaces or punctuation marks in the text to define the boundary of the sentences.

The tokens convert unstructured text into sequence of structured discrete elements, and therefore their occurrences in the document can be used directly as a structured representation of that document.

Stemming/Lemmatisation

Stemming is a process of reducing the words to their root or base form. As a result, the words are simplified and standardized, and the duplication of tokens is limited. It improves the performance of information retrieval, text classification, and other NLP tasks.

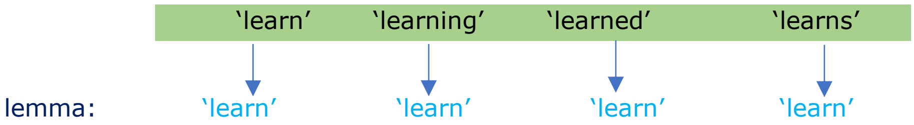
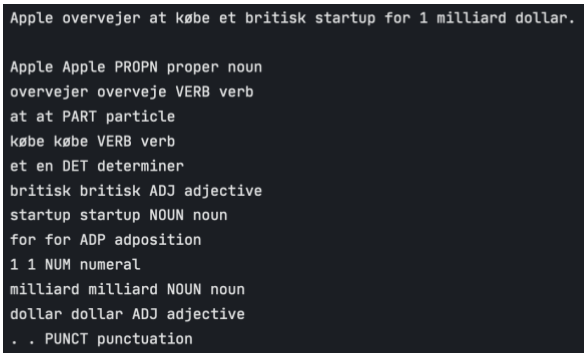


Figure 10. Stemming

POS Tagging

Part of Speech (POS) is a process of analysis and tagging the tokens with an annotation, which most likely applies to their role within the grammatical structure of sentences (such as noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, etc.



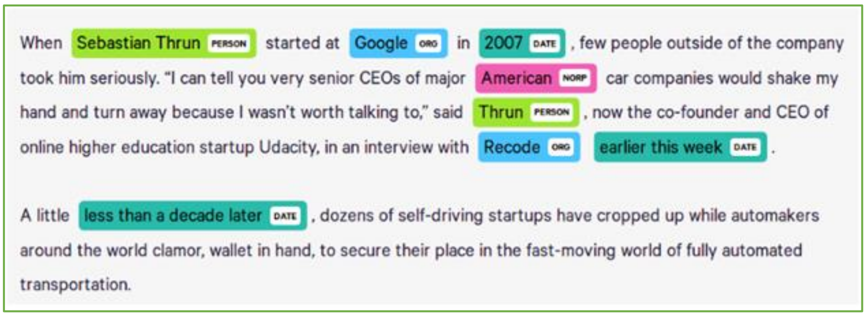
```
Apple overvejer at købe et britisk startup for 1 milliard dollar.

Apple Apple PROPN proper noun
overvejer overveje VERB verb
at at PART particle
købe købe VERB verb
et en DET determiner
britisk britisk ADJ adjective
startup startup NOUN noun
for for ADP adposition
1 1 NUM numeral
milliard milliard NOUN noun
dollar dollar ADJ adjective
. . PUNCT punctuation
```

Figure 11. POS tagging, Source

NER Tagging

Named Entity Recognition (NER) is the process of identifying and categorizing the named entities among the tokens. Such entities are names of people, organizations, locations, time, money, and currencies. An example of text tagged with NER labels is shown on Figure 12.



When Sebastian Thrun PERSON started at Google ORG in 2007 DATE , few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Figure 12. NER Categorisation Source

spaCy

The project Data Science Holodeck uses the powerful and broadly available instruments of Spacy.

2.3 Geospatial Data

Geospatial data is any type of data that enables extracting of geographic information about locations, landscapes, distances on earth. It can be a pair of geo-coordinates (latitude and longitude), post code or names of locations, and similar.

The spatial data can be in a raster format (e.g. in pixels), ready for map style visualisation, or in vector format, like a sequence of coordinates of locations and their attributes.

An example of use of geo-spatial data is shown on Figure 13.

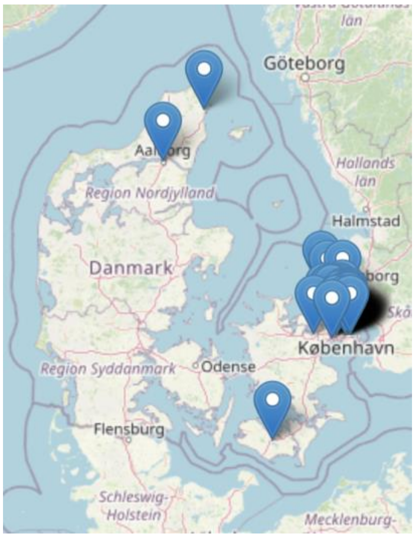


Figure 13. Geo-spatial data visualisation Source: author

The data collected from different sources usually gets ingested into data frames (tables), from where it can be explored further, and transformed into structures, appropriate for further processing.

The project Data Science Holodeck uses Folium.

Stage 3: Knowledge Graph

At the next stage we integrate all data into a domain-driven knowledge graph – a structure of nodes and edges between them. The nodes and the edges present the real-life objects, events and relations, and connected together form a full picture of the business domain area.

There are numerous advantages in introducing the graph structure in the workflow.

It is visual, intuitive, interactable, and fits well to the human mental structures.

A graph can be extended, or another way updated without affecting the already existing data. The graph dynamics resembles the model of expansion of human knowledge:

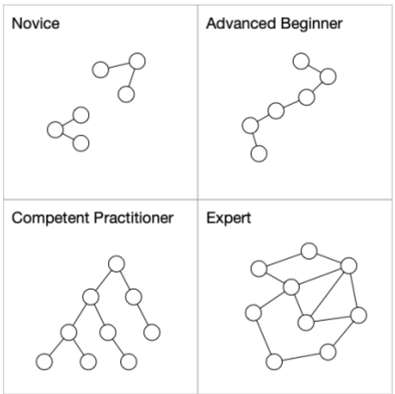


Figure 14. From beginner to expert Source

A graph is easier to explore, to search, extract patterns, or identify missing links. A knowledge graph doesn't require much of mathematics and analytical skills from the users, as most of the contained information is visible with unarmed eye. Unlike the tables, being a visual structure, the knowledge graph benefits from both the space and colouring parameters of the environment.

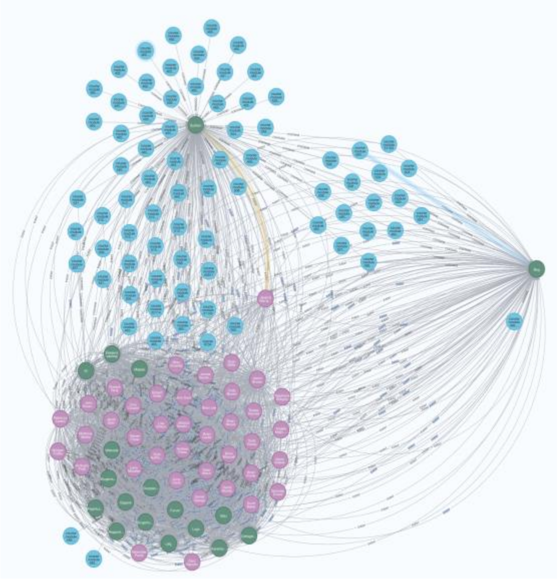
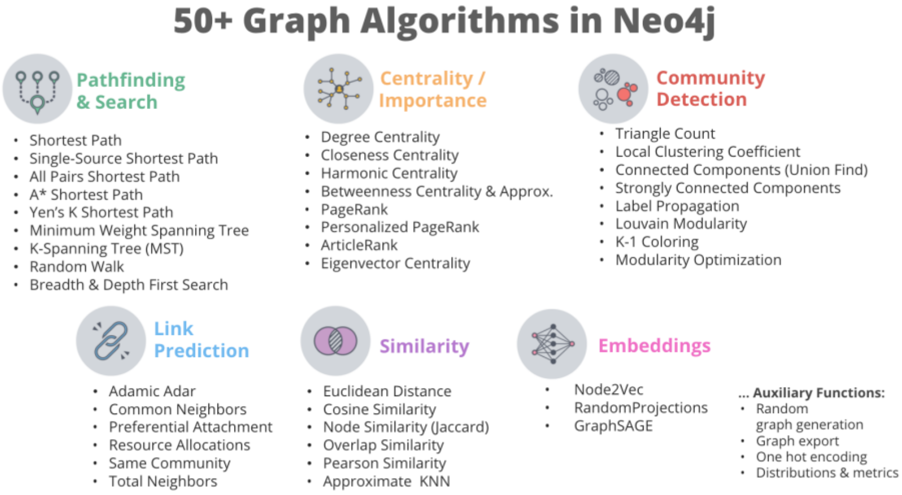


Figure 15. Knowledge graph of Moodle data in space dimensions Source: author

Storing and searching graph data are both economic and efficient in terms of volume and processing speed due to internal technical storage solutions.

The native graph algorithms, known from the graph theory add significant functionality of the models built on knowledge graphs. We are able to run operations related to topology of the graph objects and relationships, such as ranging them, finding similarity, splitting the into clusters with similar features, identifying non-typical patterns, etc. (see Figure 16).



50+ Graph Algorithms in Neo4j

Pathfinding & Search

- Shortest Path
- Single-Source Shortest Path
- All Pairs Shortest Path
- A* Shortest Path
- Yen's K Shortest Path
- Minimum Weight Spanning Tree
- K-Spanning Tree (MST)
- Random Walk
- Breadth & Depth First Search

Centrality / Importance

- Degree Centrality
- Closeness Centrality
- Harmonic Centrality
- Betweenness Centrality & Approx.
- PageRank
- Personalized PageRank
- ArticleRank
- Eigenvector Centrality

Community Detection

- Triangle Count
- Local Clustering Coefficient
- Connected Components (Union Find)
- Strongly Connected Components
- Label Propagation
- Louvain Modularity
- K-1 Coloring
- Modularity Optimization

Link Prediction

- Adamic Adar
- Common Neighbors
- Preferential Attachment
- Resource Allocations
- Same Community
- Total Neighbors

Similarity

- Euclidean Distance
- Cosine Similarity
- Node Similarity (Jaccard)
- Overlap Similarity
- Pearson Similarity
- Approximate KNN

Embeddings

- Node2Vec
- RandomProjections
- GraphSAGE

... Auxiliary Functions:

- Random graph generation
- Graph export
- One hot encoding
- Distributions & metrics

Figure 16. Categories of graph algorithms

The project Data Science Holodeck presents the knowledge graphs in a Neo4j graph database, extended with APOC and GDS (Graph Data Science) libraries.

Stage 4: AI Processing

At this stage the collected data is analysed further, with the objective of solving a particular business task.

4.1 Machine Learning (ML) and Deep Learning (DL)

Typically, there are several categories of AI processing to apply: descriptive, predictive and prescriptive analytics.

- descriptive analytics, also applied at the pre-processing stage, gathers the data and probes to define a structure and patterns, as well as discover or calculate quantifiable characteristics (for example, to compare the market behaviour per season);
- predictive analytics focuses on the historically collected data, which is used for training of data models, which can be applied for prediction of new or missing data and trends (for example, to predict the path of spreading a virus);
- prescriptive analytics simulates future scenarios and behaviour and assists in creating plans and recommending actions (for example, recommending investments or another positive user experience).

The NLP and NLU applications involve all of the approaches mentioned above. NLP/NLU is a domain of artificial intelligence aimed at understanding, interpreting, and generating natural language. It can include a description of the non-structured documents, prediction or prescription of new sources, or full generation of new documents and missing links.

The methods and algorithms, applied to the process can be described as supervised and non-supervised regression, classification and clustering methods from ML and DL.

4.2 Natural Language Processing and Understanding (NLP/NLU)

Typical NLP and NLU tasks the project solve are key words extraction, extractive and abstractive summarization, and questions answering.

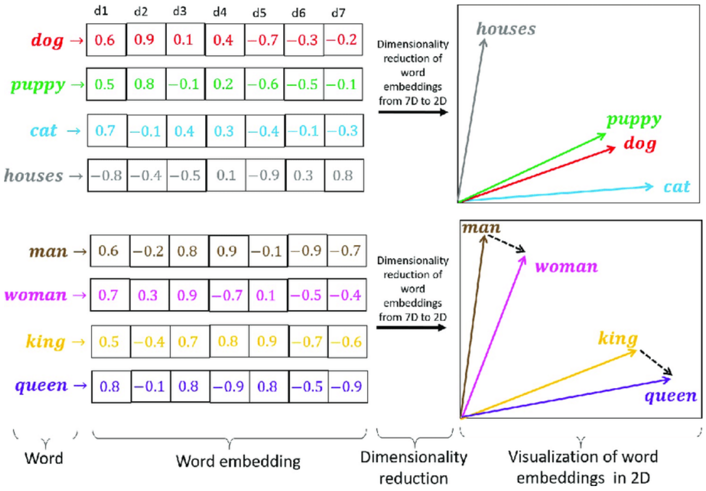
The related operations require the pre-processing techniques, defined at the previous stage: tokenization, POS, and NER.

Vectorisation

The next necessary step in closing the gap between the text and the processing algorithms is the digitalisation / vectorisation of the text – transformation of the tokens into vectors of numbers. The operation is called embedding. The numbers identify the text components, but also capture their relationships.

The numbers in the vectors can also be seen as coordinates of points, located on a map. Just like locations of objects shown on geographical map tells us how close these locations are in reality, the numbers if any two embedding vectors tell how close these vectors are, and therefore, how similar the semantic meanings if the vectorised texts are. Two expressions close in meaning would be represented by two similar vectors of numbers (see Figure 17).

We are able to embed not only text, but images, graph objects, or other databases' entities.



	d1	d2	d3	d4	d5	d6	d7
dog →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
puppy →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
cat →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Dimensionality reduction of word embeddings from 7D to 2D
houses
puppy
dog
cat
Dimensionality reduction of word embeddings from 7D to 2D
man
woman
king
queen
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Figure 17. Word Embedding Source

Vector Databases

The vector embeddings are stored in a vector database, where from they can be searched, extracted, and compared for similarity (see Figure 17).

The vector databases are relatively new and emerging type of databases, especially used in AI applications. They are designed, scaled, and optimised for efficient storage and fast retrieving of vector data in vast volumes. The data is encoded by applying intelligent algorithms and the access to the data is facilitated by advanced indexing systems, enabling rapid approximate similarity search.

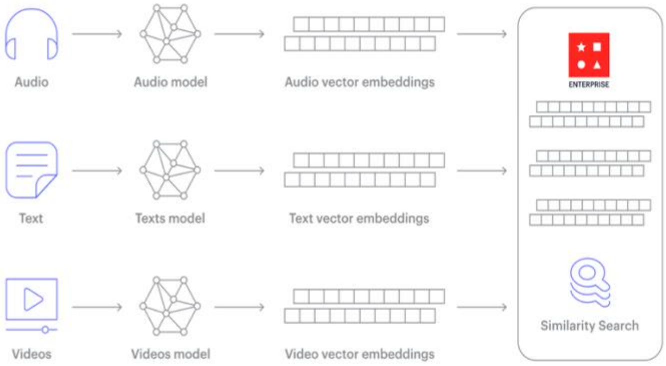


Figure 17. Vector Database Source: Redis

The project Data Science Holodeck applies several embedding instruments from GDS, BERT, and LangChain platforms.

The recent revolutionary implementations of NLP/NLU introduce new methods and algorithms, built over the popular categories of the Artificial Neural Networks (ANN) – convolutional, recursive, and recurrent neural networks.

Attention Mechanism

The attention mechanism is a ML technique for improving the performance of the training ANN by enabling them to identify and select the most relevant parts of the input sequences It achieves it by assigning different weights to the sentence embeddings, giving higher value to the most relevant. Attention is used in solving tasks as machine translation and text summarisation, to mention some.

Self-Attention

A mechanism applied in NLP to a sequence (sentence) for the purpose of computing its representation by relating selected words in that sequence. The self-attention avoids the limitation of processing the whole sequence in one and the same order. The self-attention allows variety of internal combinations between the sub-sequences to find the most significant ones.

Both attention and self-attentions have been defined and presented in the paper "Attention Is All You Need", Vaswani et al. (2017), and have become some of the most efficient instruments in the modern AI.

Transformers

The title *transformer* is given to a specific architecture of Neural Network efficiently solving sequence-to-sequence language tasks (Figure 18). The main innovation of the transformer is the self-attention mechanism, mentioned above, which allows training a model, while paying attention to different parts of the input sequence simultaneously. All traditional neural networks process the entire input sequence sequentially, which creates difficulties in capturing the dependencies between the parts.

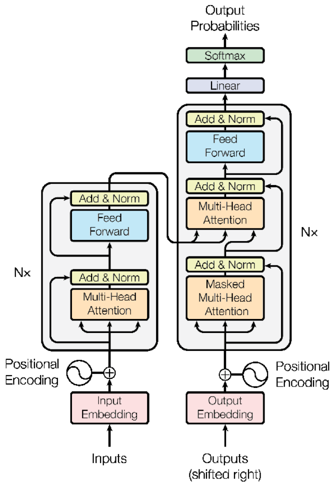


Figure 18. Transformers Source

The attention and transformer stay behind the success of ChatGPT.

Large Language Models (LLM)

LLM are transformer-based NLP neural networks, which try to predict the text that would come next in the sequence. The models are trained by observing huge number of parameters.

There are two groups of LLM, trained in either supervised mode (BERT - Bidirectional Encoder Representations) or in unsupervised (GPT - Generative Pre-trained Transformer).

LLMs are powerful implementation of transformers and attention in variety of NLP/NLU projects, such as text summarization, text generation, sentiment analysis, recommendation systems, fraud detection, content creation, chatbots, virtual assistants, and conversational AI, speech recognition and synthesis, machine translation.

The project Data Science Holodeck applies a selection of LLMs hosted by the platforms Hugging Face and Lang Chain.

Stage 5: 3D and VR visualisation

Nowadays, it is not sufficient to provide the consumers with relevant results, it is better to provide meaningful knowledge based on these results and allow them using it effectively.

Active knowledge is more powerful gear than static information. We can activate the knowledge by interacting with the data, modelling and simulating various situations and observe the behaviour of the system and the data structure.

5.1 3D Visualisation

Interactive 3D visualisation (Figure 19) is more efficient in understanding the scenes than the 2D diagrams and classic dashboards. It brings advantages, such as:

- enables running the application on a desktop or mobile display, without a need of additional equipment
- intuitive, user friendly, and efficient
- doesn’t require skills in operating the complex dashboards provided by the BI vendors

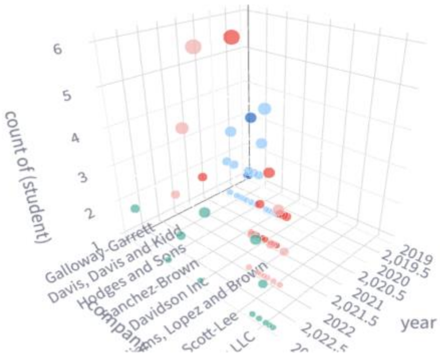


Figure 19 3D Interactive3D visualisation

The 3D interactive visualisation doesn’t serve well in all scenarios. When the data volumes are very large or exploration tasks are complex, the VR visualisation serves the users and the applications much better.

5.2 VR Visualisation

VR adds value providing immersive experience. It opens a completely different perspective of viewing the data, actively involving the user in ‘communication’ with the data. It enables humans to step into a virtual world built for solving a particular domain-specific task, seeing both the big picture of it and highlighting its components, freely changing the viewpoints.

VR data promotes attention to details and facilitates variety of cognitive activities, such as comparison, selection, pattern recognition, etc. People can get deep insights in the matter, without proficiency in data analytics or graph theory, as no other environment can enable. It can cut off the distance between the human intuition and the abstract numbers and enlighten the relationships between the abstract concepts.

Inside the VR environment, people can use either controller devices or hand gestures to interact with the data objects, make updates, and observe the effect. Dynamically changing the scenarios and parameters enables collecting vast information in short time, helping the vision of data-driven predictions and the effect of knowledge-driven decisions.

Both the organisations and the learners in higher education can gain from bridging the gap between the algorithms and their practical implementation, between seeing and acting, between one-directional to bi-directional analysis, and therefore accelerating the performance and the outcomes of the whole task.

The project Data Science Holodeck applied 3D and VR visualisation solutions in two ways:

- built as Unity application with XR Interaction Toolkit

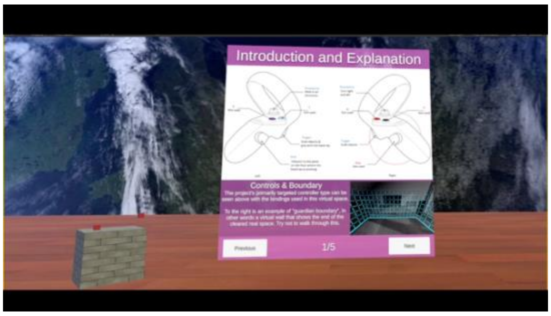


Figure 20. Interactive VR Data Visualisation

- built as an interactive web application, where the concept of force-directed graphs and the libraries ThreeJS, A-Frame, and d3-force-3d are implemented.

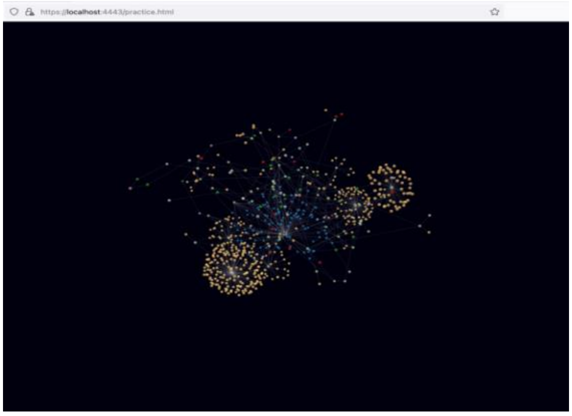


Figure 21. Interactive 3D and VR Graph Visualisation

For more examples and detailed explanation, see the Demo Cases at the project's web site.

5.3 Deployment

We have created several applications, demonstrating the use and variations of the suggested technology.

Demo Case 1 (Figure 20) is an Android application build in Unity C# for running in Oculus and Oculus 2 VR headsets.

Demo Case 2 (Figure 21) shows a web-based Python-built client-server application deployed by means of Streamlit. It can run on either local or remote server, displaying the visualisations on both computer screen and VR headset's display by means a standard web browser.