PAGODE

Semantic enrichment using natural language processing and entity extraction methods

Vassilis Tzouvaras

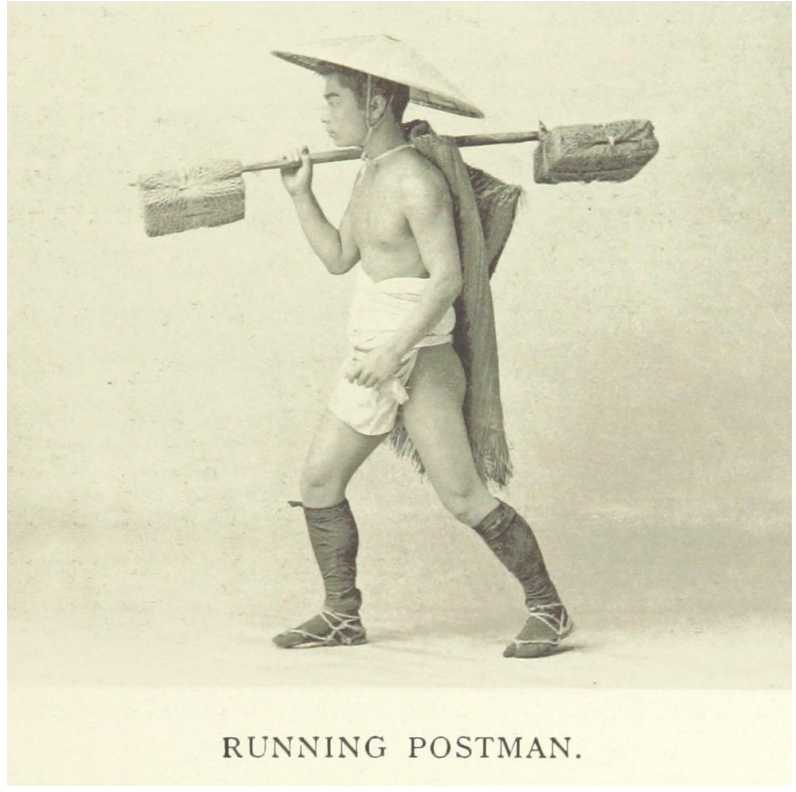
made with europeana

DIGITAL MEETS CULTURE
www.digitalmeetsculture.net

Metadata Enrichment

Why is needed?

- Quality of Metadata
- XML, RDF, LoD Vocabularies, URIs, SKOS, WikiData, Geonames
- Manual process
- Automatic enrichment, machine learning
- Crowdsourcing, Human in the loop,

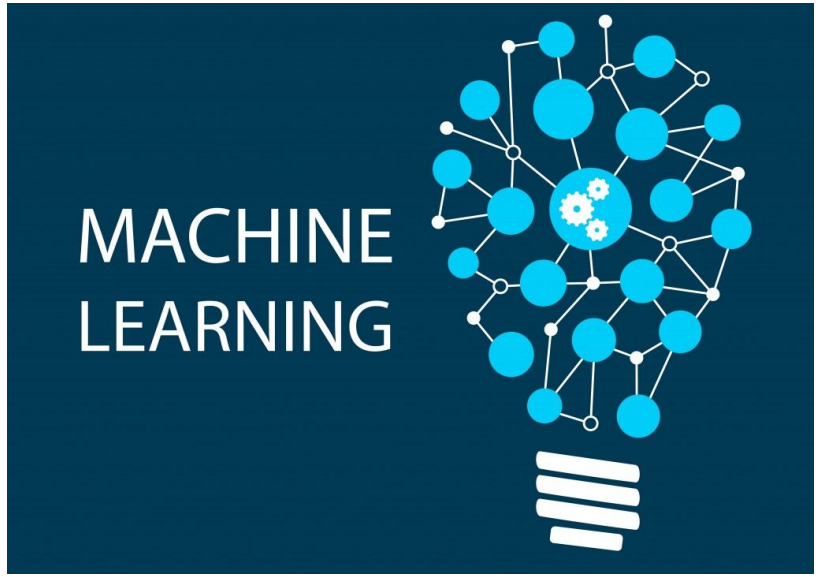


Letter carrier from "Letters from the Land of the Rising Sun".1886 - 1892, British Library
United Kingdom, Public Domain

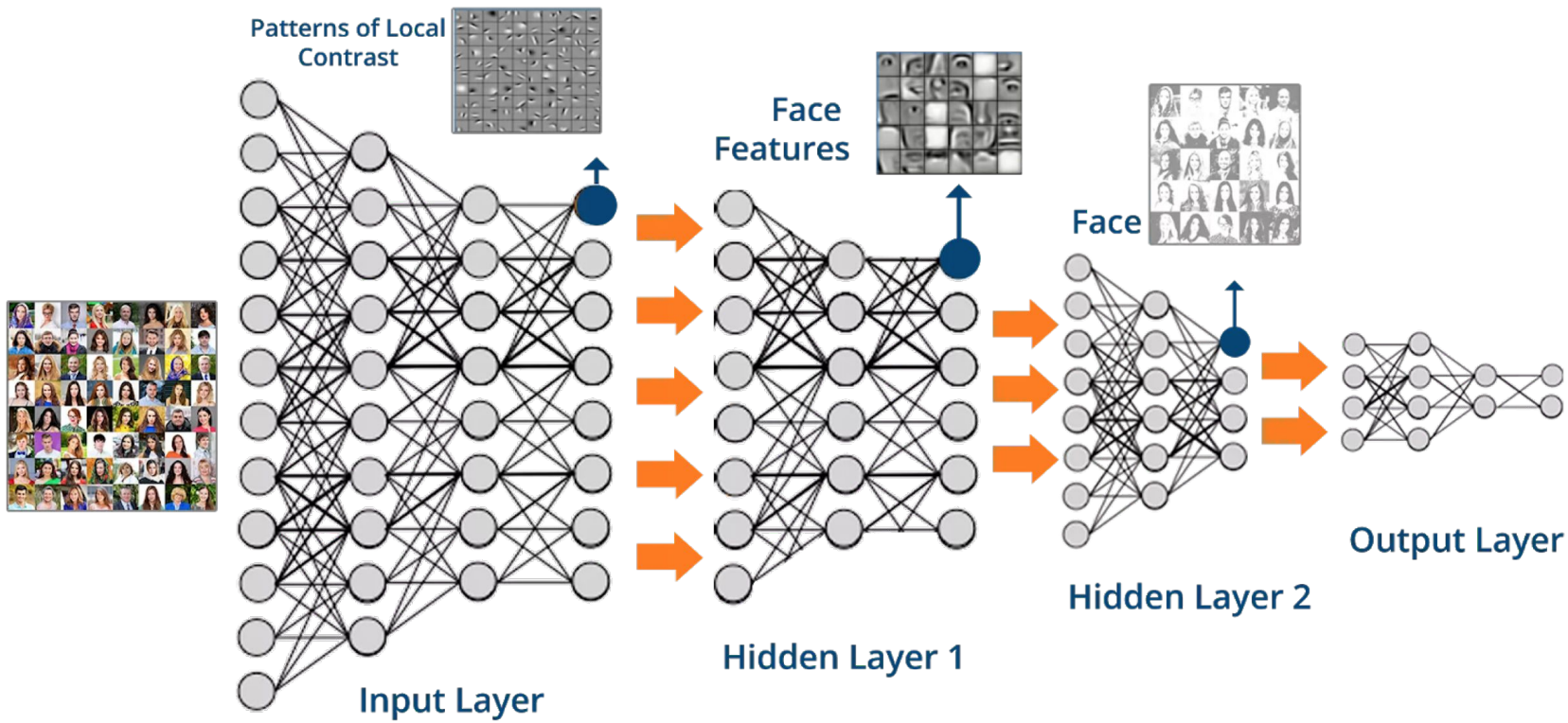
Machine Learning

Machine learning algorithms build a mathematical model of sample data, known as "**training data**", in order to make predictions or decisions

Technologies: NLP, object detection, machine translation, event detection, photo aesthetics, visual similarity,

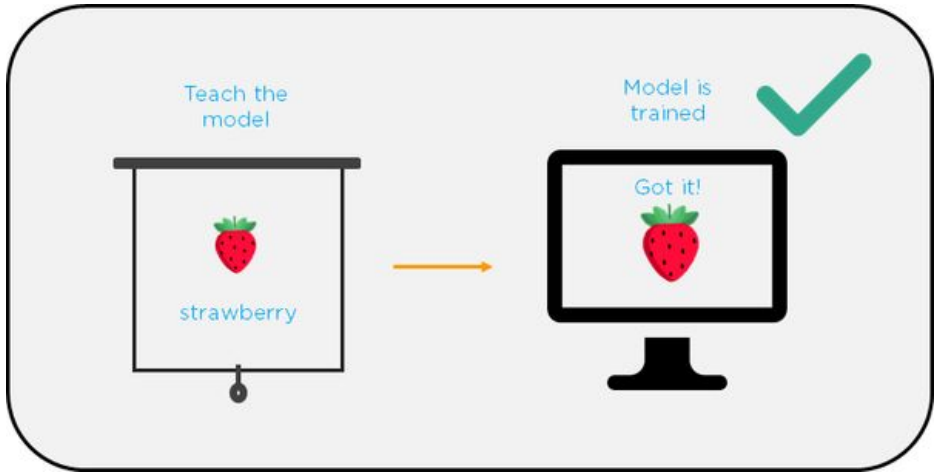


Machine Learning through Neural Networks



Supervised Learning

Supervised learning algorithms build a mathematical model of a set of data that contains both the **inputs** and the **desired outputs**. The data is known as **training data**, and consists of a set of training examples.



Datasets???

Machine learning requires high volumes of data for training, validation, and testing.

it’s important to have the right data, structured in the right format, covering all the variation of your solution.

So how do you get the large volume of structured data you need?
Human-annotated data is the key to successful machine learning.

Crowd & Machine Intelligence

Machine intelligence and **human** intelligence can cooperate and improve each other in a mutually rewarding way.

- Exploit the user obtained annotations for training/improving machine learning algorithms
- Use machine learning methods to validate user acquired labels
- Crowdsourcing campaign with specifically selected content which will improve performance of automated machine learning system

Semantic Enrichment in Pagode

- Enrich metadata that is already in Europeana
- Use Europeana's items API to extract items in the form of datasets
- Enrich datasets and ingest back to Europeana through the Annotations API
- Use of Community (list of terms transformed in SKOS thesaurus) and international Vocabularies (Wikidata, Geonames)
- Semantic Enrichment tool built by NTUA in Europeana XX CEF project
- Validation using the Crowdheritage platform built by NTUA in Crowdhertiage CEF project

Challenges

- Machine learning is not always producing correct results
- Validation and crowd intelligence
- Lack of datasets to train networks
- Use of pre-trained networks or networks that use unsupervised learning
- Build a user friendly enrichment platform (for expert users)
- Use appropriate vocabularies
- Rich text
- multilinguality