Vergangenheit, Gegenwart und Zukunft biblischer Handschriftenentschlüsselung: Small Data Analysis - KI und das Potential der Entschlüsselung antiker Texte und Artefakte

Sophie Robert-Hayek & Annette Weissenrieder

University of Lorraine
Martin-Luther-University Halle-Wittenberg

Introduction to computational humanities

Definition

Digital Humanities

Computational Humanities are:

- an interdisciplinary field;
- combining research in traditional humanities;
- with **tools from computer science and mathematics**;
- to bring new knowledge to humanities related problems.

Definition

Digital Humanities

Computational Humanities are:

- an interdisciplinary field;
- combining research in traditional humanities;
- with **tools from computer science and mathematics**;
- to bring new knowledge to humanities related problems.

Computational humanities bring together experts from a wide range of disciplines:

- **humanities**
- **mathematicians**
- **computer scientists**

to provide new answers and perspectives on existing problems.

What Are Digital Humanities?

Recent advances in computing offer **unprecedented opportunities** to

- **Generate**

What Are Digital Humanities?

Recent advances in computing offer **unprecedented opportunities** to

- **Generate**
- **Explore**

What Are Digital Humanities?

Recent advances in computing offer **unprecedented opportunities** to

- **Generate**
- **Explore**
- **Interpret**

data.

The integration of computing with humanities disciplines promises **new perspectives for research, analysis, and understanding of existing data**.

Introduction to Machine Learning approaches

The AI revolution

The last 10 years have seen a synergy between:

- The definition of new algorithms;
- The ability to store and to collect data;
- The augmentation of computing speed.

All these factors have led to the birth of a new research field: Data science and Machine Learning

Machine Learning and Deep Learning

Machine Learning algorithms

Algorithms able to **learn** and **adapt** without following explicit instructions by **drawing inferences from patterns in data**.

Machine Learning and Deep Learning

Machine Learning algorithms

Algorithms able to **learn** and **adapt** without following explicit instructions by **drawing inferences from patterns in data**.

Given a **training** dataset, Machine Learning algorithms are able to **find patterns in data** to **predict** or **infer** information on new data.

Machine Learning and Deep Learning

Machine Learning algorithms

Algorithms able to **learn** and **adapt** without following explicit instructions by **drawing inferences from patterns in data**.

Given a **training** dataset, Machine Learning algorithms are able to **find patterns in data** to **predict** or **infer** information on new data.

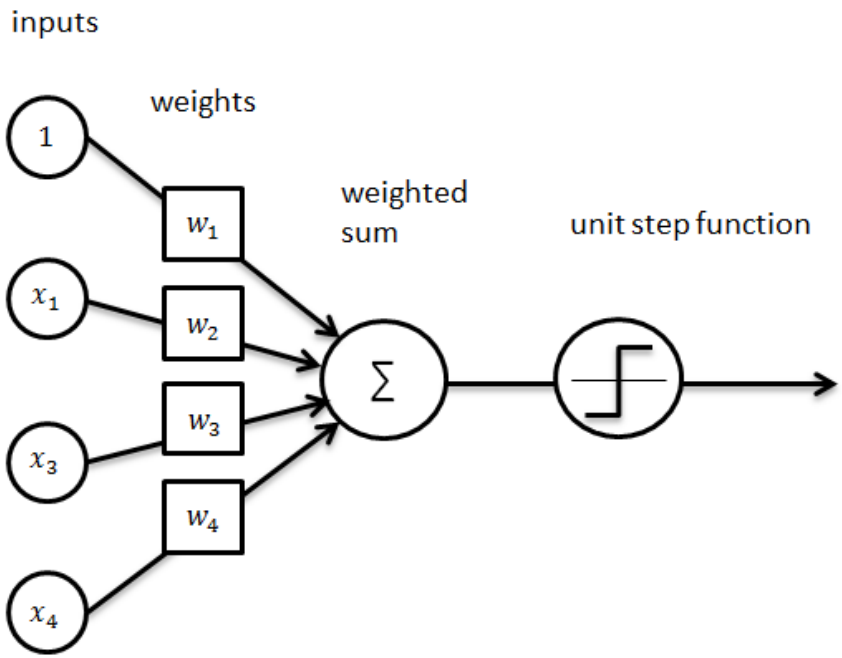
Deep Learning

Deep learning consists in using a sub-class of machine learning methods, called **artificial neural networks**.

Deep Learning principles

Perceptron

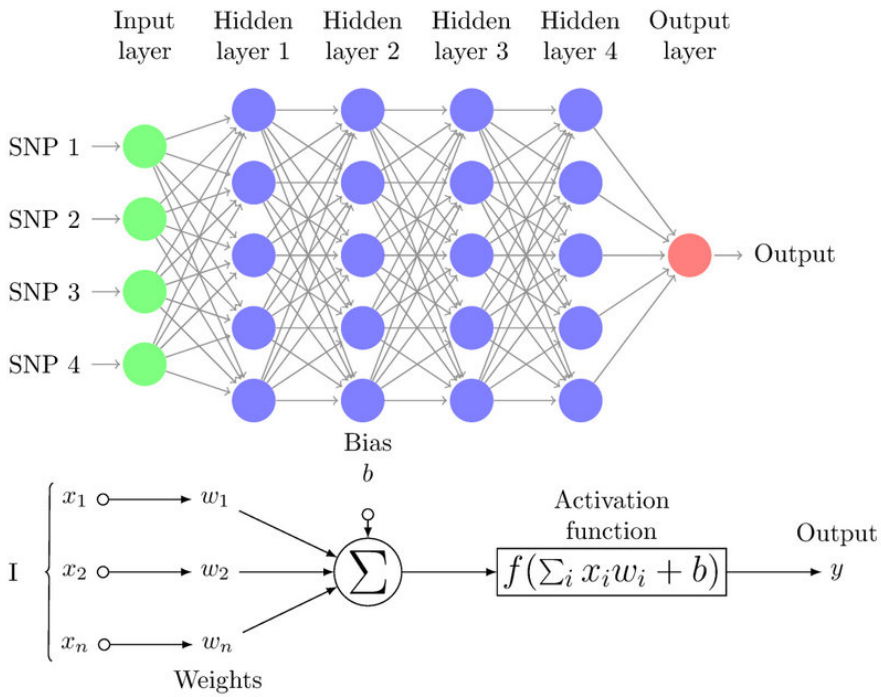
A **perceptron** is a type of **binary** linear classifier, as a model of how the human processes information. It is the building block of more complex neural network architectures.



Deep Learning principles

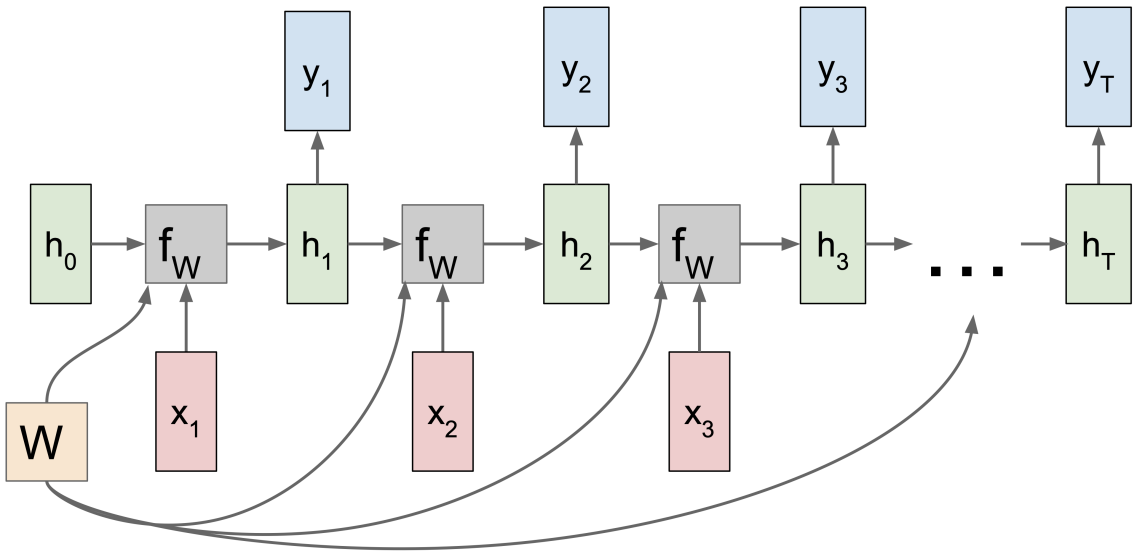
Multi-layer perceptrons

The **multi-layer perceptron (MLP)** is formed by stacking multiple single-layer perceptrons together, making it **deep**. It becomes composed of an **input layer**, **hidden layers** and **output layers**.



Sequence based neural networks

The current explosion in the performance of Machine Learning is due to new kind of neural network architecture **able to process sequences**: LSTM and Transformers.



Major fields of application

Because these neural networks are able to process sequences, they are very performant in the fields of:

Natural Language Processing

Natural Language Processing (NLP) aims to enable computers to **comprehend, interpret, and interact with human language in a manner that is both meaningful and contextually appropriate.**

Major fields of application

Because these neural networks are able to process sequences, they are very performant in the fields of:

Natural Language Processing

Natural Language Processing (NLP) aims to enable computers to **comprehend, interpret, and interact with human language in a manner that is both meaningful and contextually appropriate**.

Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text.

Machine Learning, Deep Learning and philology

Applying these novel techniques to **philology** and **manuscript studies** is called **computational philology**

Two very exciting perspectives:

- **Optical Character Recognition**: ability to perform large scale numerization for the study of the materiality of manuscripts and their content.
- **Natural Language Processing**: ability to **process large amount of texts** and **find patterns otherwise undected**;

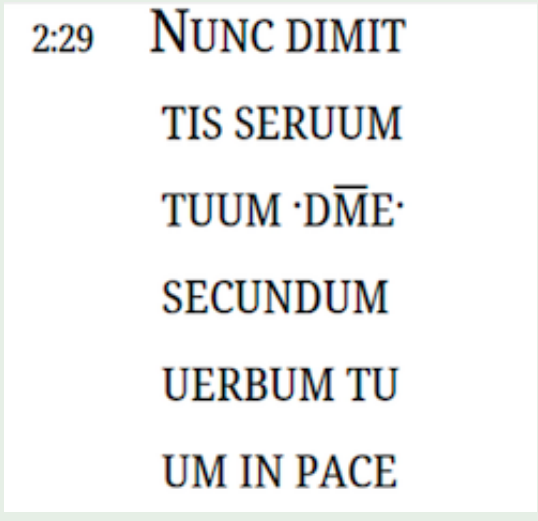
Natural Language Processing and Codex Vercellensis

Automatic approaches for the study of text

Natural Language Processing neural networks allow to quickly:

- Perform lemmatization;
- Perform stylistic analysis of texts;
- Compute distances between manuscripts;
- Perform alignment of words between different languages.

Automatic lemmatization



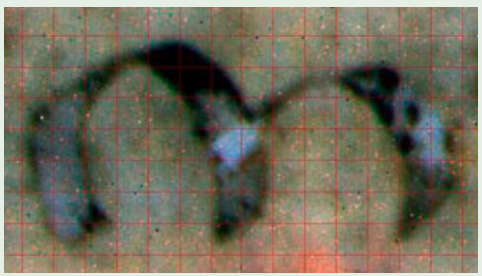

```
Word nunc -> nunc
Word dimittis -> dimitto
Word seruum -> seruus
Word tuum -> tuus
Word domine -> dominus
Word secundum -> secundum
Word uerbum -> uerbum
Word tuum -> tuus
Word in -> in
Word pace -> pax
```

Study of paleography and style

Possibility to perform very quickly **the computation of different stylistic attributes of Vercellensis**:

- Vocabulary used;
- Morphological analysis of the words used;
- Grammatical constructions of the text;
- Count occurences of specific paleographic form.

Paelographic study of letter shape



Very quickly count the frequency of apparition of a certain morphology of M.

Study of word morphology and grammatical style

Study the writing style of manuscript by performing the analysis of the word form:

```
Word nunc ->
=== POS: ADV
Word dimittis -> Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin|Voice=Act
=== POS: VERB
Word seruum -> Case=Acc|Gender=Masc|Number=Sing
=== POS: NOUN
Word tuum -> Case=Acc|Gender=Masc|Number=Sing
=== POS: DET
Word domine -> Case=Voc|Gender=Masc|Number=Sing
=== POS: NOUN
Word secundum ->
=== POS: ADP
Word uerbum -> Case=Acc|Gender=Neut|Number=Sing
=== POS: NOUN
Word tuum -> Case=Acc|Gender=Neut|Number=Sing
=== POS: DET
Word in ->
=== POS: ADP
Word pace -> Case=Abl|Gender=Fem|Number=Sing
=== POS: NOUN
```

Study of inter-textuality

Possiblity to automatically:

- **compute vocabulary and stylistic distances between different manuscripts**;
- better understand the relationship between the different latin texts through **automatic analysis and collations**.

Bezae: *nunc dismittis serbum tuum domine secundum uerbum tuum in pace*
Vercellensis: *nunc dimittis seruum tuum domine secundum uerbum tuum in pace*.

```
      Bezae   nunc   dismittis serbum   tuum domine secundum uerbum tuum in pace
Vercellensis  nunc    dimittis seruum   tuum domine secundum uerbum tuum in pace

Number of word differences: 2
```

Study of renderings and translation techniques

Possibility to perform automatic alignment between ancient languages to better understand:

- Relationship between manuscripts of different languages (Syriac, Greek);
- Translation techniques.

Vercellensis	**Vaticanus**
nunc	Νῦν
dimittis	ἀπολύεις
seruum tuum	τὸν δοῦλόν σου
domine	δέσποτα
secundum	κατὰ
uerbum tuum	τὸ ῥῆμά σου
in pace	ἐν εἰρήνῃ

Computer vision and Codex Vercellensis

Computer vision and the study of the Vercellensis

Hubert Mara