The Whys and Hows in "I Trust AI"
Objectives, methods, expected outcomes

Luciana Duranti and Muhammad Abdul-Mageed
University of British Columbia
San Benedetto del Tronto, 7 July 2023

InterPARES TrustAI

Authenticity in Digital Records

The assessment of the **authenticity of digital material**

- **is always an inference** based on their significant properties, as shown by their identity metadata, and
- **relies on circumstantial evidence** such as
  - the **integrity of the system** hosting it at any given time,
  - the **policies and procedures controlling its life**,
  - the **technology encrypting or securing the access to it,** and
  - the **persons responsible for keeping or preserving it**

Keeping and preserving records so that they maintain **authenticity through time** is even harder

InterPARES TrustAI

Can Artificial Intelligence Help?

We know that Artificial Intelligence Systems provide

- Inconclusive Evidence (based on probabilities)
- Inscrutable Evidence (no interpretability or transparency)
- Misguided Evidence (as good as the data provided)
- Unfair Outcomes (disproportionate impact on some groups of people)
- Transformative Effects (challenges for autonomy and privacy)
- Non Traceability (hard to assign responsibility)

Plus

- The decisions AIS make are based on past decisions, and
- when it comes to human affairs, tomorrow rarely resembles today, and data can't say what has a moral value, nor what is socially desirable

InterPARES TrustAI

Montreal Declaration Principles (2018)
Canada and the Association of Southeast Asian Nations

- Respect for Well-being principle
- Respect for Autonomy Principle
- Protection of Privacy Principle
- Solidarity Principle
- Democratic Participation Principle
- Equity Principle
- Diversity and Inclusion Principle
- Caution Principle
- Responsibility Principle
- Sustainable Development Principle

InterPARES TrustAI

Past Experience with AI

There have been several projects looking at Artificial Intelligence for controlling and accessing records: they typically look at a particular tool in a specific context or even a single set of records.

- recurrent neural networks for classification of the content of large aggregations of records
- recommendation systems that connect relevant images to digitized letters, by using handwritten text recognition (HTR) to make digitized documents searchable
- chatbots that emulate human conversation through voice commands or text chats or both to help knowledge seekers find connected information
- a combination of Named Entity Recognition (NER), entity relations tools, and topic modeling to create visualization tools for the types of data stored on disk images

InterPARES TrustAI

The Problem for Archives

- Relying on existing off the shelf tools, as all the studies on AI for records have done, limits what challenges can be met, as it makes the needs of record governance subservient to the larger field of machine learning
- It may be practical, but many tangible instances of bias have been found in modern machine learning models, often driven by questionable data collection practices
- This raises the questions of a) whether off the shelf tools are the best solution for records and b) what AI could look like if this power relationship between AI and archives were reversed, with **archival theory informing the creation of AI tools.** Our answer has been to begin another phase of InterPARES.

InterPARES TrustAI

InterPARES

- The **InterPARES** (International research on Permanent Authentic Records in Electronic System) project, funded by the Social Sciences and Humanities Research Council of Canada, **has addressed digital records issues since 1998**, focusing on **current and emerging technologies** as they evolve, and developing **theory, methods, and frameworks** that allow for the **ongoing trusted preservation of the records resulting from the use of such technologies**.
- Its 5th and latest iteration, *I Trust AI*, differs from the previous ones as it is not concerned with the records produced by a specific technology, but has the purpose of **using AI to carry out archival functions for the control in the long term of all records, on any medium, and from any age**, and to do so in such a way that the **trustworthiness** of the records remains protected and verifiable, and that the tools and processes are **transparent, unbiased, equitable, inclusive, responsible, and sustainable**

InterPARES TrustAI

I Trust AI Project Goal

The overall goal of the latest phase of the InterPARES research project, I Trust AI, is to design, develop, and leverage Artificial Intelligence to support the ongoing availability and accessibility of trustworthy public records by forming a sustainable, ongoing partnership producing original research, training students and other highly qualified personnel (HQP), and generating a virtuous circle between academia, archival institutions, government records professionals, and industry, a feedback loop reinforcing the knowledge and capabilities of each party.

InterPARES TrustAI

Objectives

- Identify specific AI technologies that can address critical records challenges;
- Determine the benefits and risks of using AI technologies on records;
- Ensure that records concepts and principles inform the development of responsible AI; and
- Validate outcomes from Objective 3 through case studies and demonstrations.

The more than 40 **case studies and general studies** in course relate to several aspects of records creation and retention and disposition, preservation and access. They are carried out by more than 200 researchers (31 countries and 92 organizations) organised in international and multidisciplinary teams (including experts from AI, records and archives management, forensics and law, data science, engineering, etc.).

InterPARES TrustAI

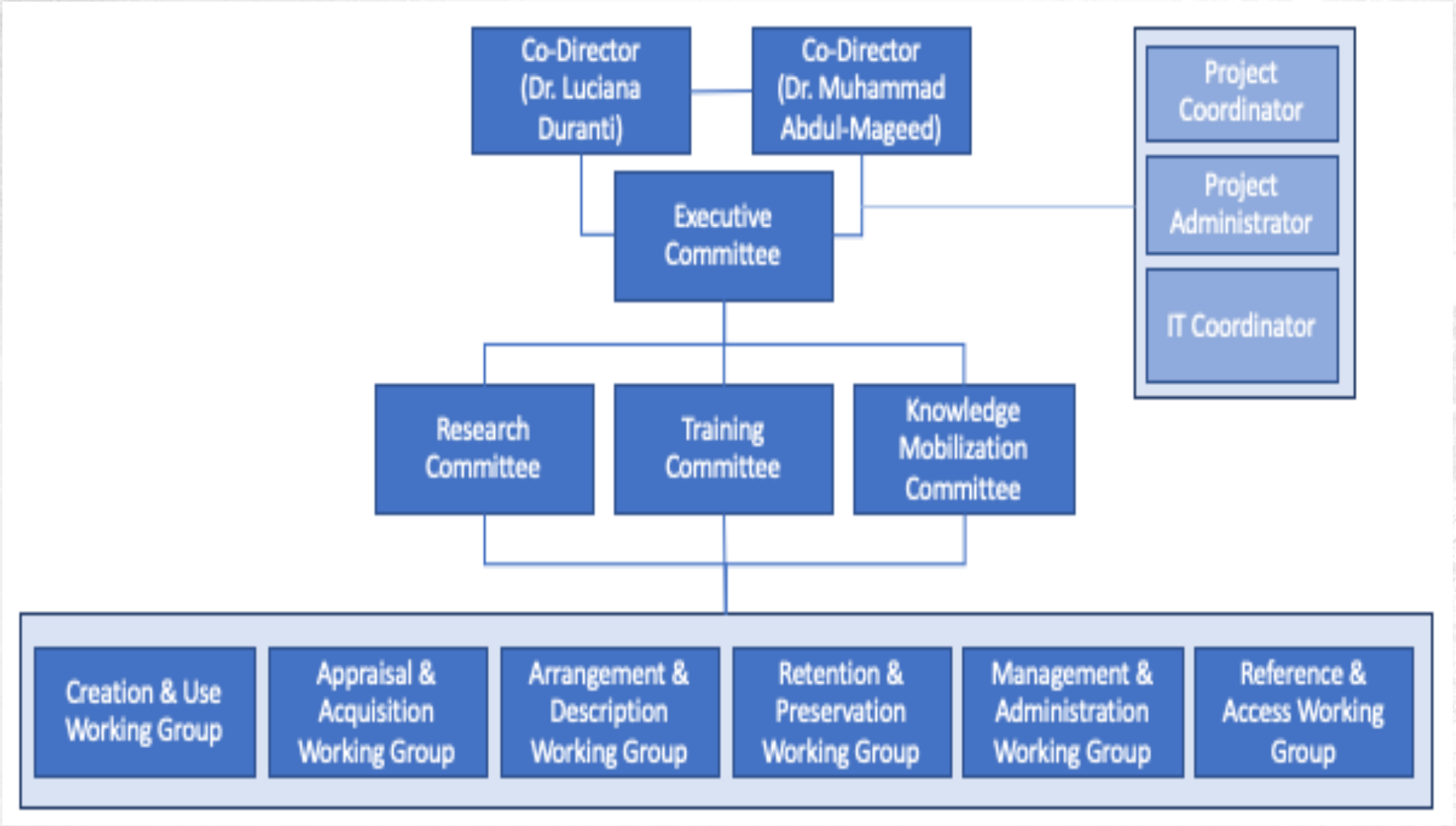
Expected Outcomes

The project will improve upon existing tools and create new Machine Learning tools that will address records needs, such as

- machine translation,
- image recognition and description,
- optical character recognition (OCR) and handwritten text recognition,
- text summarization and classification, and
- text style transfer for language civilization (e.g., removal of bias, hate, and sexism)

InterPARES TrustAI

Organization of the project



InterPARES TrustAI

Studies

- Studies are **all international and interdisciplinary**
- Focus on all aspects of archival functions
  1. Creation and use of trustworthy records
  2. Appraisal and acquisition of archival material
  3. Arrangement and description
  4. Retention and preservation
  5. Management and administration of records and archives
  6. Reference and access

InterPA
Trus

Methods

Our approach is **two-pronged**, comprising the practical and immediate need to address large-scale existing problems, and the longer-term need to have AI-based tools that are reliably applicable to future problems.

- Our short-term approach focuses on **identifying high impact problems and limitations in records and archives functions, and applying AI to improve the situation**. This will be achieved via collaboration between records and archival scientists and professionals and AI researchers and industry experts.
- Our long-term approach focuses on **identifying the tools that records and archives specialists will need in the future to flexibly address their ever-changing needs.** This includes decision support and, once decisions are made, rapid implementation of AI-based solutions to those needs.

InterPARES TrustAI

Methods (cont.)

- The fact that the *I Trust AI* project is a **multinational interdisciplinary endeavour** means that our first effort must be to **understand each other, starting with the terms we use**. For example, archival professionals talk about **records**, while computer scientists and AI professionals talk about **data**. To archivists, data are the smallest meaningful unit of information in a record. To an AI specialist, data are organized information (possibly in a database), be it facts or not, regardless of size, nature and form.
- Thus, key to our work are **AI tutorials and workshops** for non-AI researchers, and **archival and diplomatics theory tutorials** for non archival researchers. These educational endeavours are supported by the **Terminology Database** which is developed in collaboration by a multidisciplinary team.
- It is clear to every InterPARES researcher that the dominant perspective is archival because of the goal and objectives of the project, so archival terminology is to be used in the project outputs.

InterPARES TrustAI

Methods (cont.)
AI for InterPARES purposes

- We use **Deep learning (DL), a sub-field of machine learning (ML), which are both sub-fields of AI**.
- DL mimics information processing in the brain. This is possible by designing artificial neural networks arranged in multiple layers that **take input, attempt to learn a good representation of it, and map it to an output decision**.
- DL methods learn best when given large amounts of **labeled data** (e.g., for a model that detects sensitive information, labels can be from the set *sensitive, not-sensitive*). One of our case studies looks at the identification of Personally Identifiable Information in records. We aim to label data according to **diplomatic methodology**, starting with the identification of persons (e.g. author, addressee) in each type of record. This type of learning with labeled data is called ***supervised learning***. There are also *semi-supervised* and *unsupervised learning* used in various case studies.

InterPARES TrustAI

The Whys and Hows in "I Trust AI": Objectives, Methods, and Expected Outcomes

Muhammad Abdul-Mageed
The University of British Columbia
Twitter: @mageed

San Benedetto del Tronto (2023-07-07)

I Trust AI: The Archives Question

- **AI systems** for **records and archives that maintain the *nature* and *trustworthiness* of the records**

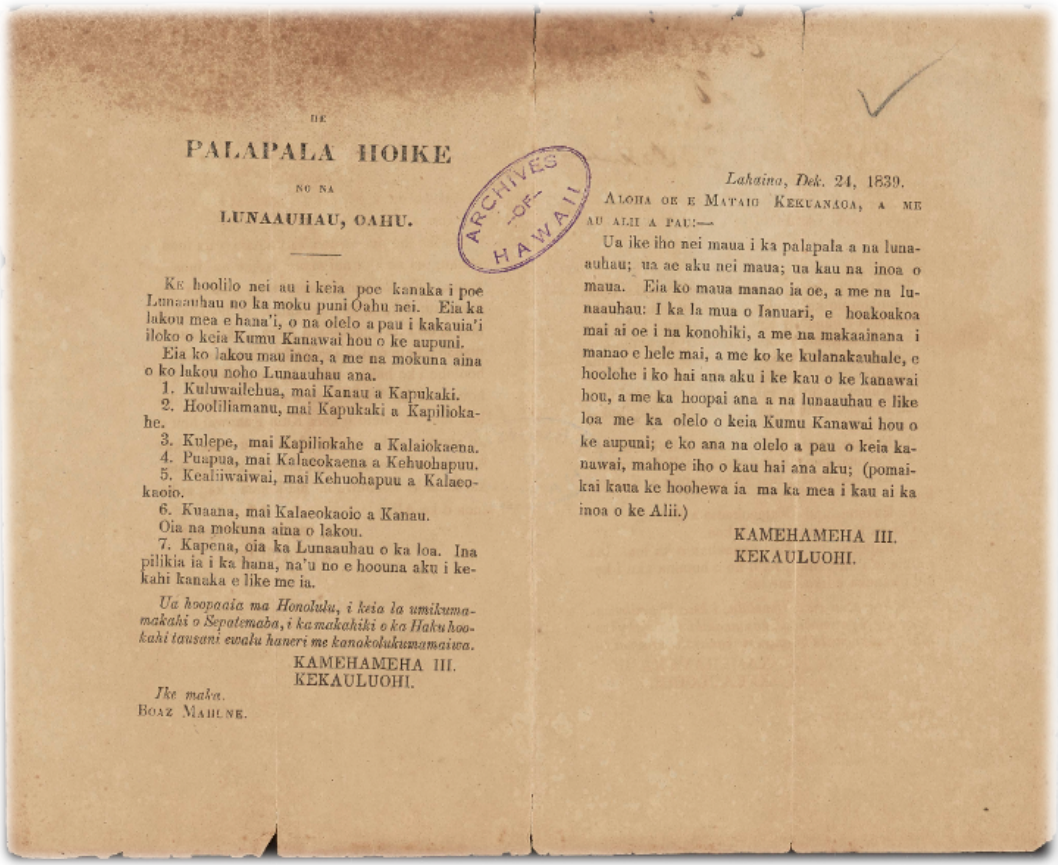
Archival Functions & AI Tasks

- Map archival functions into **AI tasks**
- Good tasks: clearly defined
- Develop **AI methods** for tasks
- Focus: **SoTA, satisfy the needs,**

  . . .

Example: Access → OCR

Challenge: Other languages (e.g., Hawaii Langs)



HE

PALAPALA HOIKE

NO NA

LUNAAUHAU, OAHU.

KE hoolilo nei au i keia poe kanaka i poe Lunaauhau no ka moku puni Oahu nei. Eia ka lakou mea e hana'i, o na olelo a pau i kakauia'i iloko o keia Kumu Kanawai hou o ke aupuni.

Eia ko lakou mau inoa, a me na mokuna aina o ko lakou noho Lunaauhau ana.

1. Kuluwailehua, mai Kanau a Kapukaki.
2. Hooliliamanu, mai Kapukaki a Kapiliokahe.
3. Kulepe, mai Kapiliokahe a Kalaiokaena.
4. Puapua, mai Kalaeokaena a Kehuohapuu.
5. Kealiiwaiwai, mai Kehuohapuu a Kalaeokaoio.
6. Kuaana, mai Kalaeokaoio a Kanau.

Oia na mokuna aina o lakou.

7. Kapena, oia ka Lunaauhau o ka loa. Ina pilikia ia i ka hana, na'u no e hoouna aku i kekahi kanaka e like me ia.

Ua hoopaaia ma Honolulu, i keia la umikumamakahi o Sepatemaba, i kamakahiki o ka Hakuhookahi tausani ewalu haneri me kanakolukumamaiwa.

KAMEHAMEHA III.
KEKAUAULUOHI.

Ike maka.
BOAZ MAHUNE.

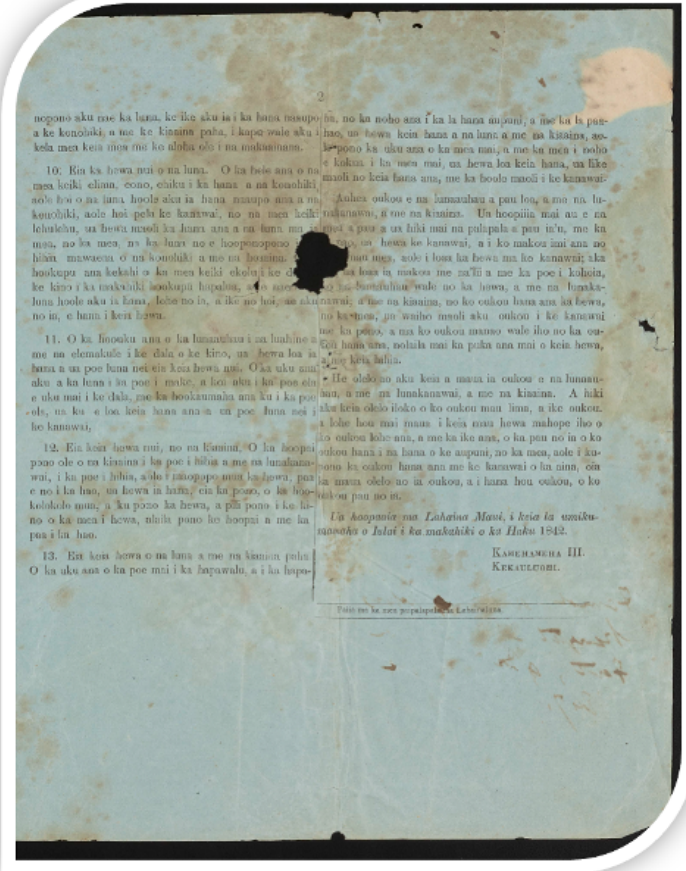
ARCHIVES OF HAWAII

Lahaina, Dek. 24, 1839.

ALOHA OE E MATAIO KEKUANAOA, A ME AU ALII A PAU:—

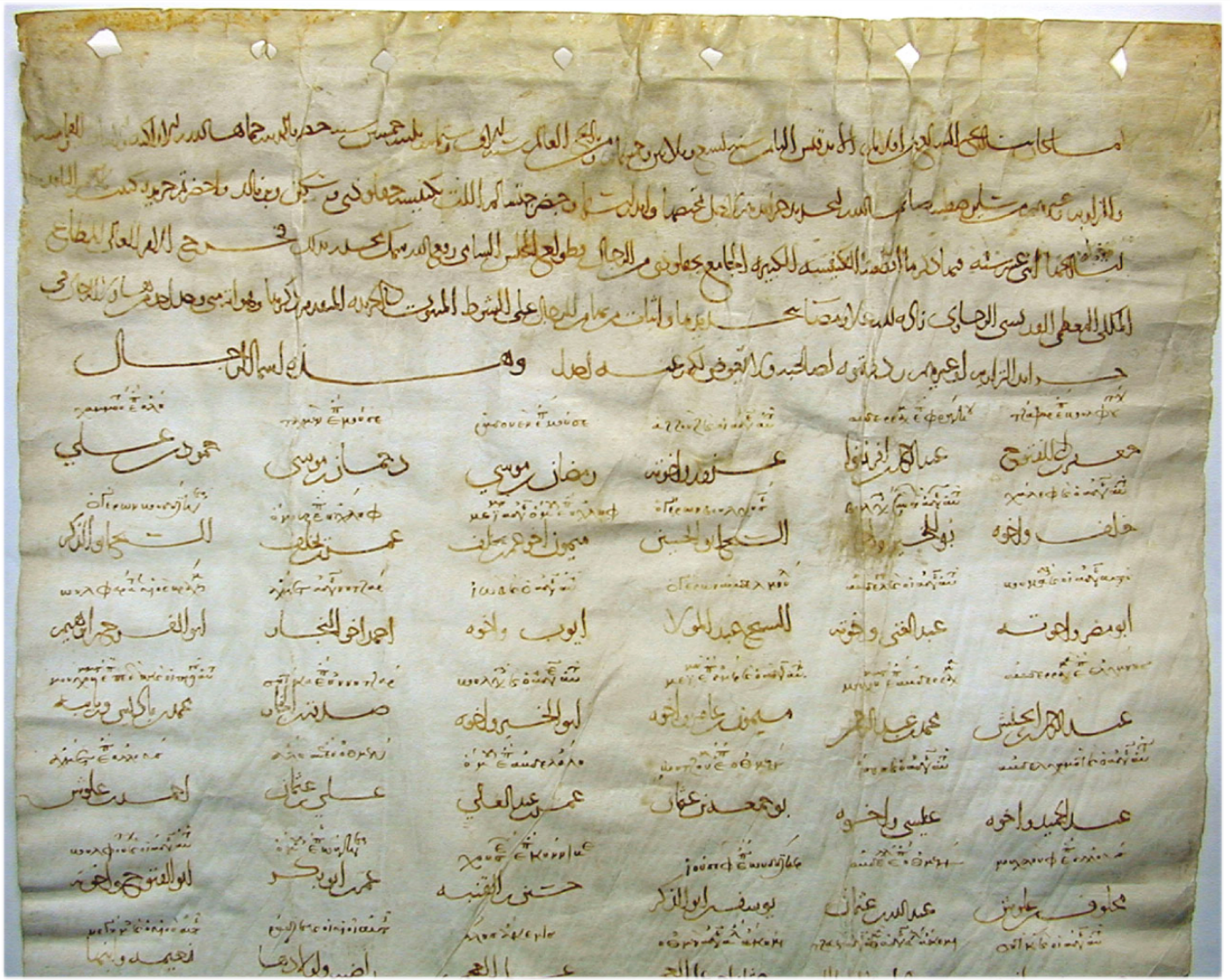
Ua ike iho nei maua i ka palapala a na lunaauhau; ua ae aku nei maua; ua kau na inoa o maua. Eia ko maua manao ia oe, a me na lunaauhau: I ka la mua o Ianuari, e hoakoakoa mai ai oe i na konohiki, a me na makaainana i manao e hele mai, a me ko ke kulanakauhale, e hoolohe i ko hai ana aku i ke kau o ke kanawai hou, a me ka hoopai ana a na lunaauhau e like loa me ka olelo o keia Kumu Kanawai hou o ke aupuni; e ko ana na olelo a pau o keia kanawai, mahope iho o kau hai ana aku; (pomaikai kaua ke hoohewa ia ma ka mea i kau ai ka inoa o ke Alii.)

KAMEHAMEHA III.
KEKAULUOHI.



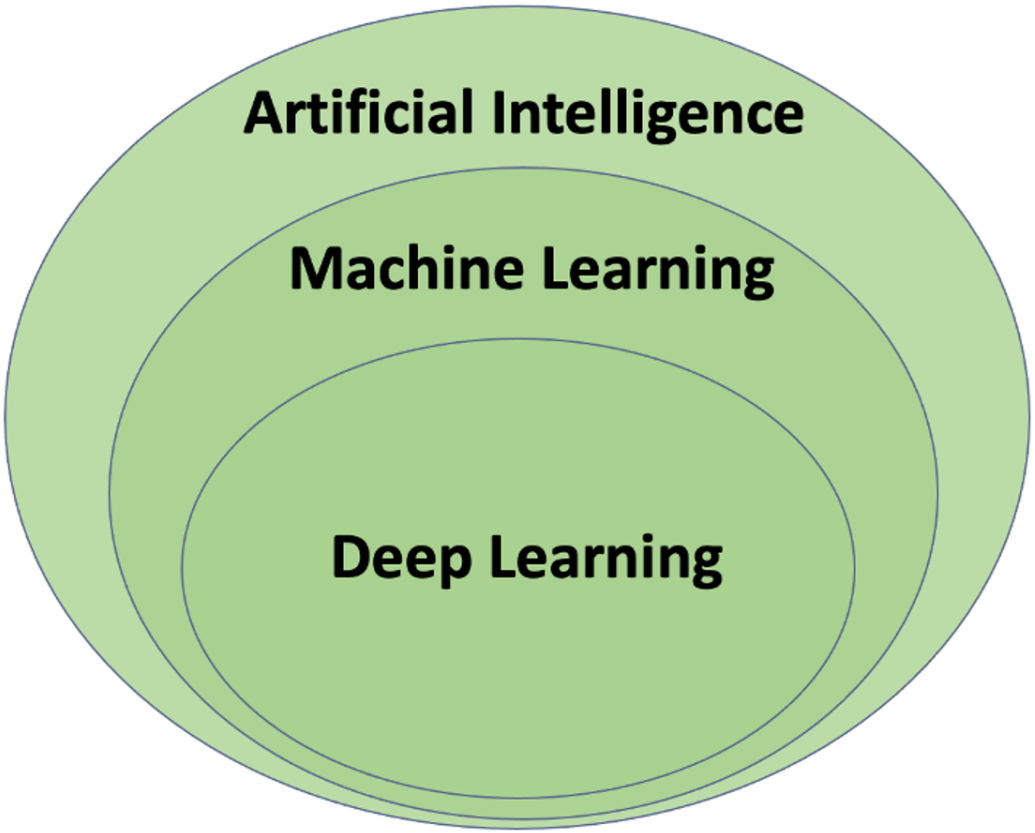
Challenge: Non-Latin, HWR

Challenge: Multilingual Parchments

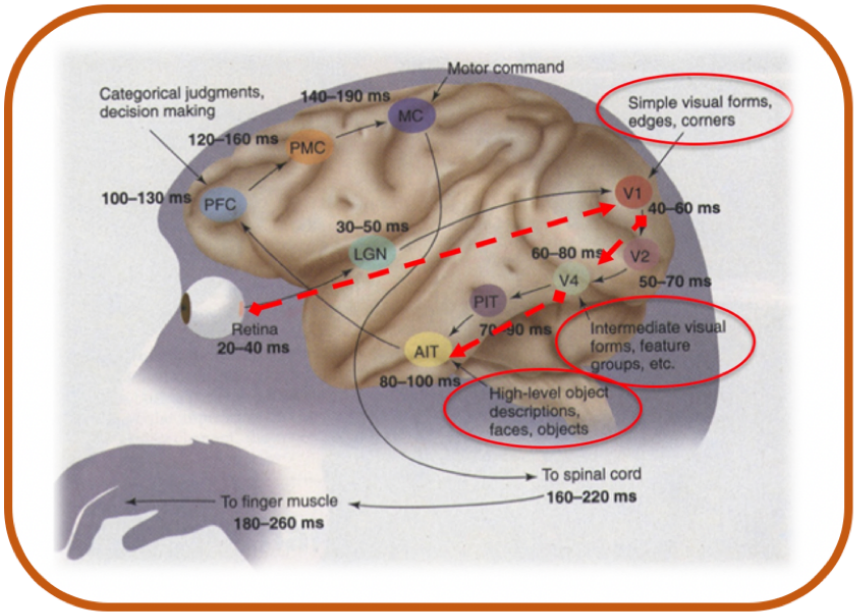


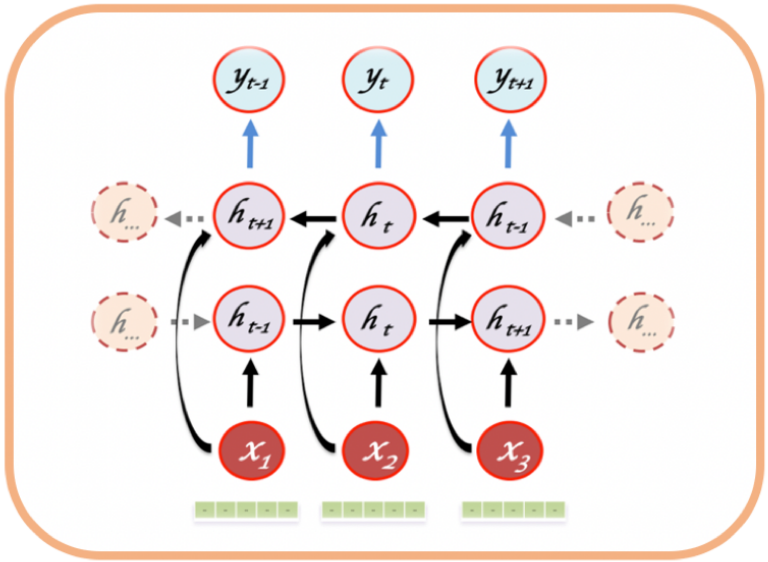
Working with Data

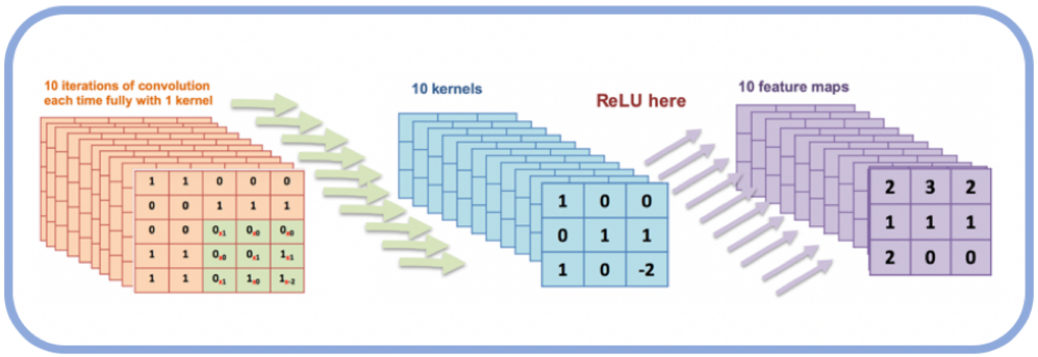
Artificial Intelligence

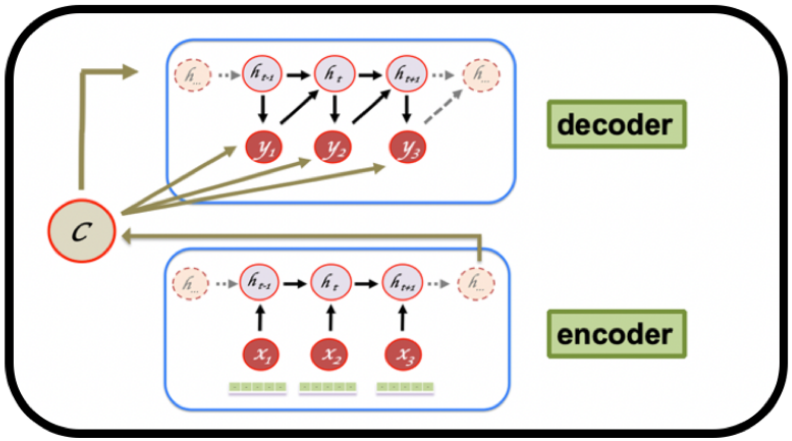


Deep Learning



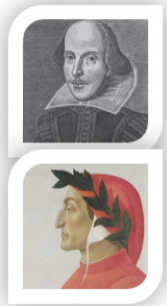




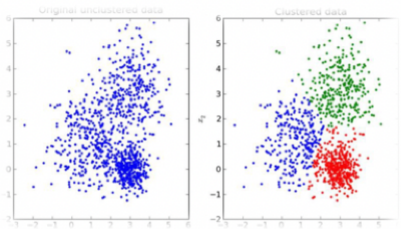


Machine Learning & Supervision

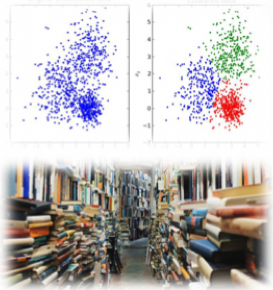
- **Supervised**

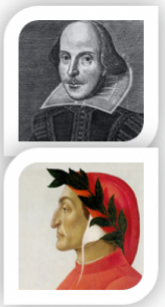


- **Unsupervised**

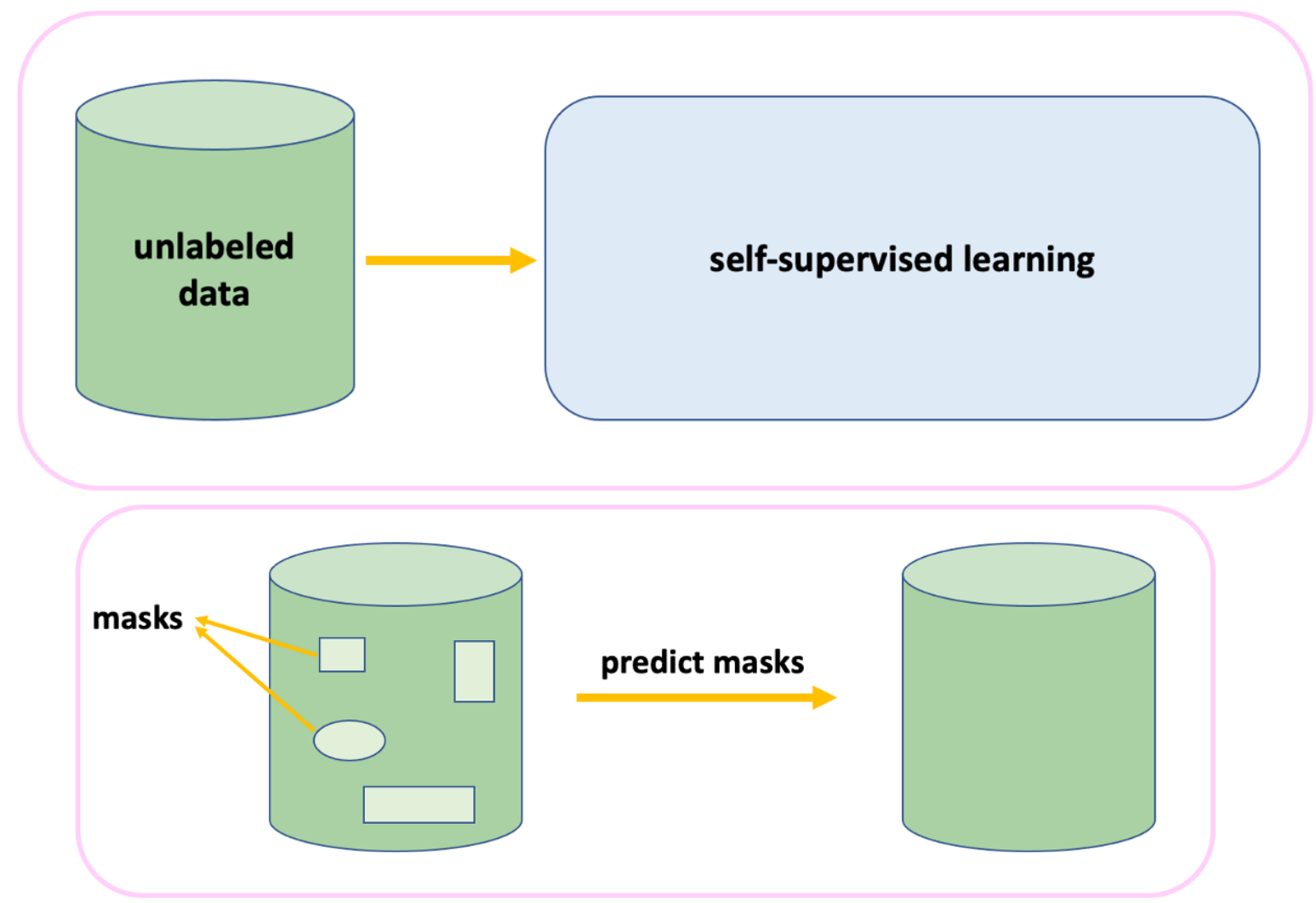


- **Semi-supervised**

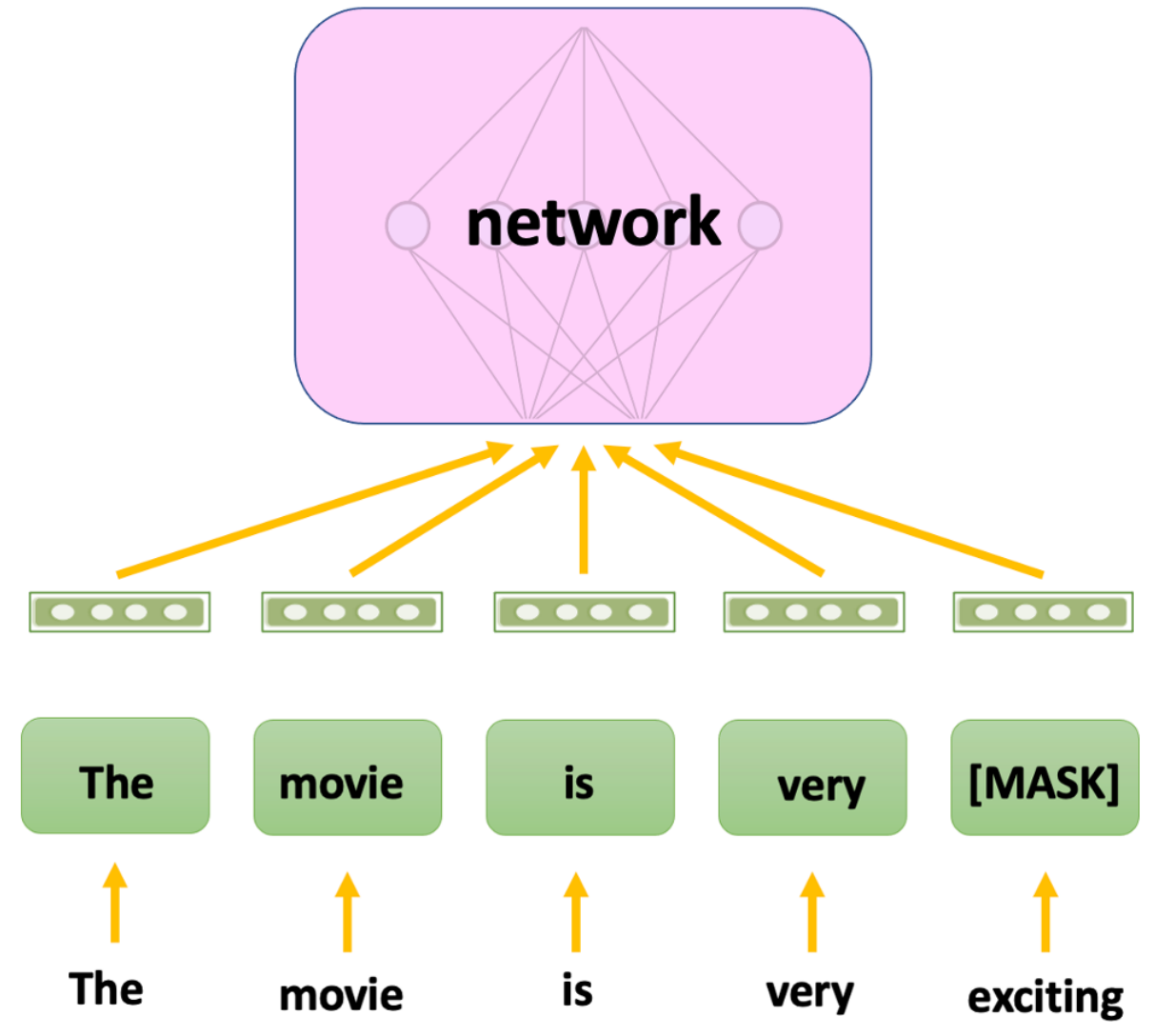




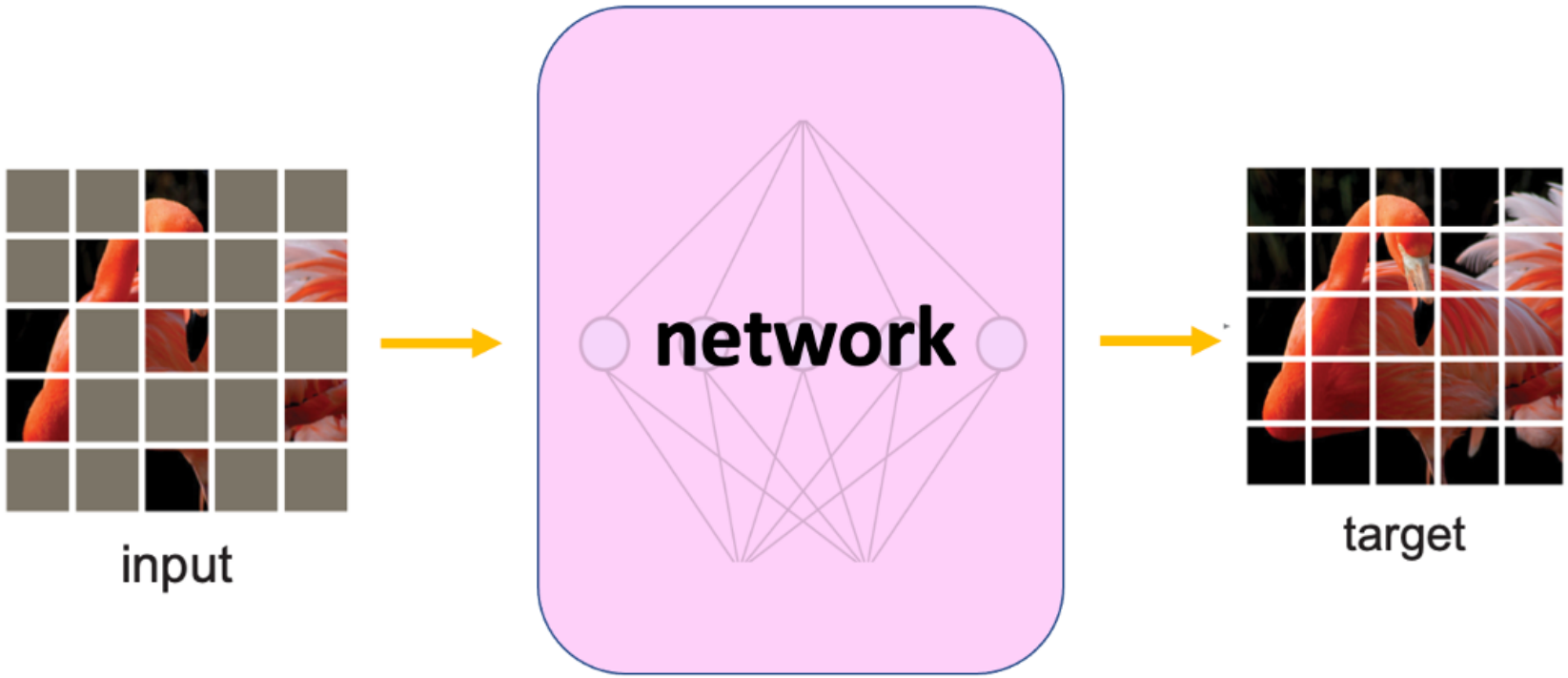
Self-Supervised Learning



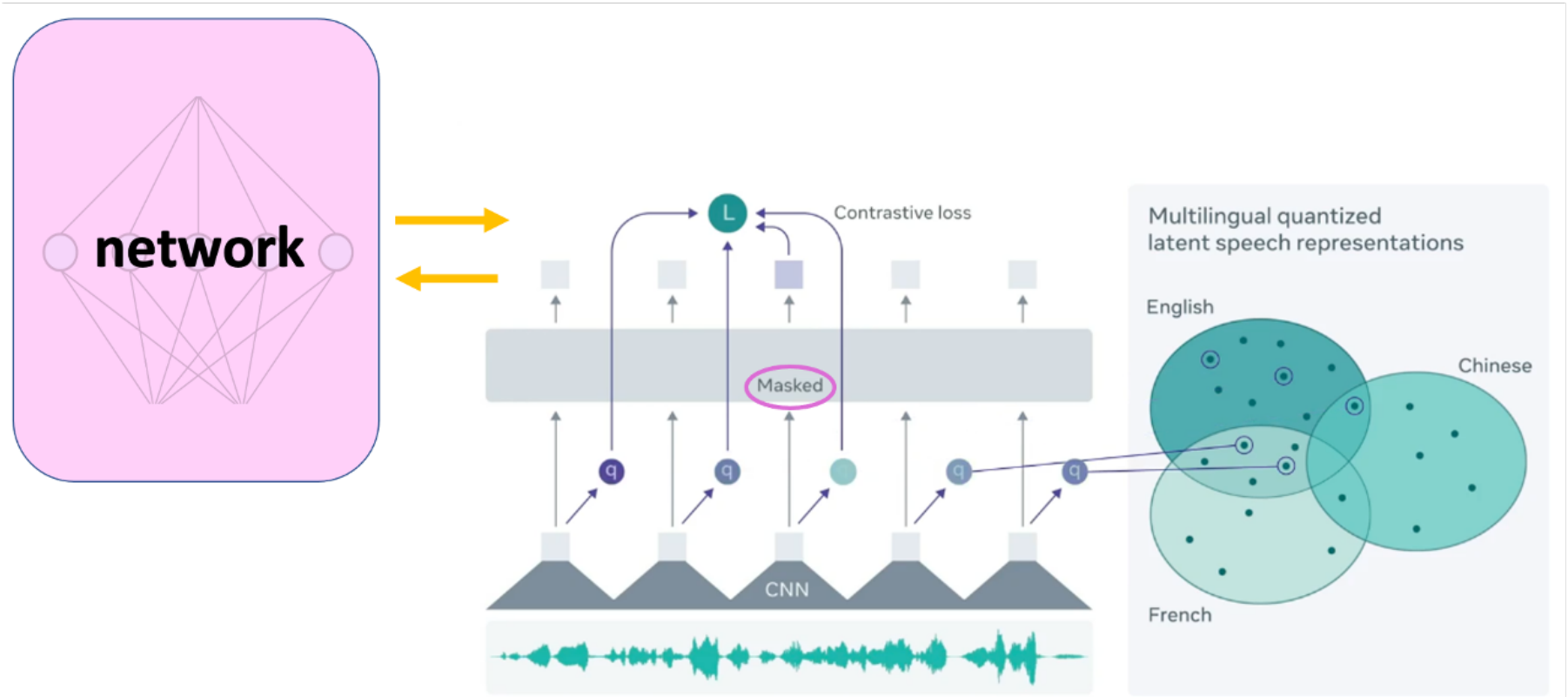
Text



Image



Speech


network
L
Contrastive loss
Masked
q
CNN
Multilingual quantized
latent speech representations
English
Chinese
French

Denoising Diffusion Models

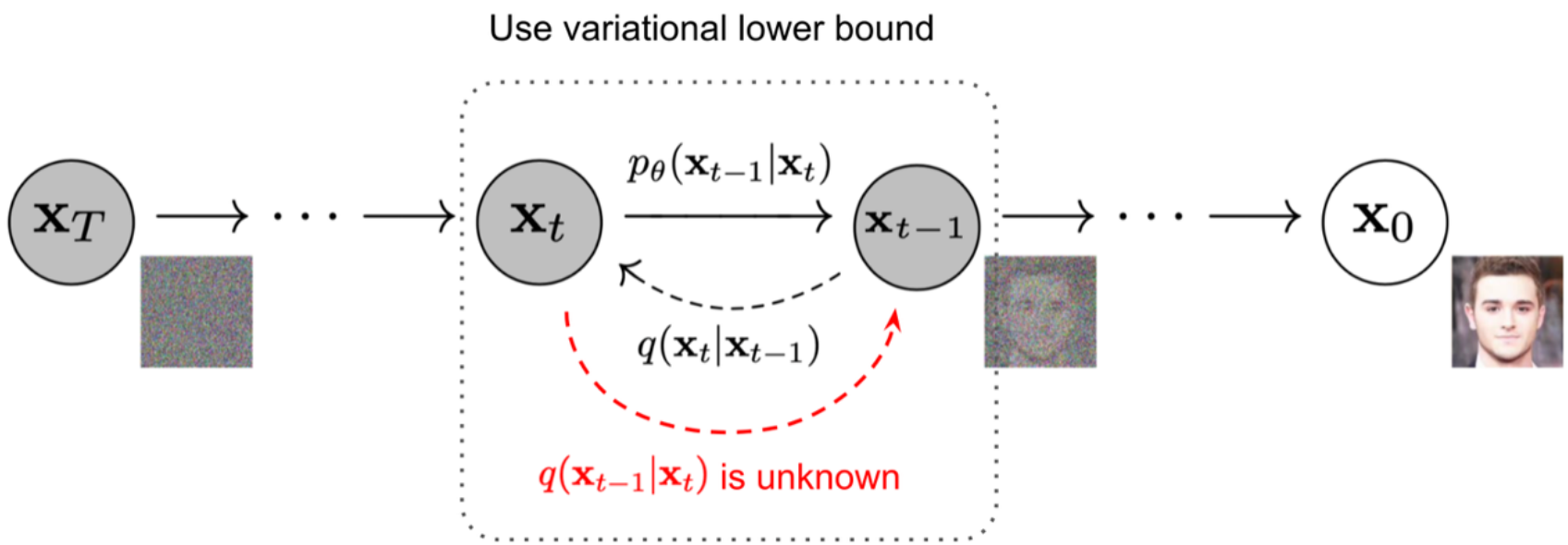
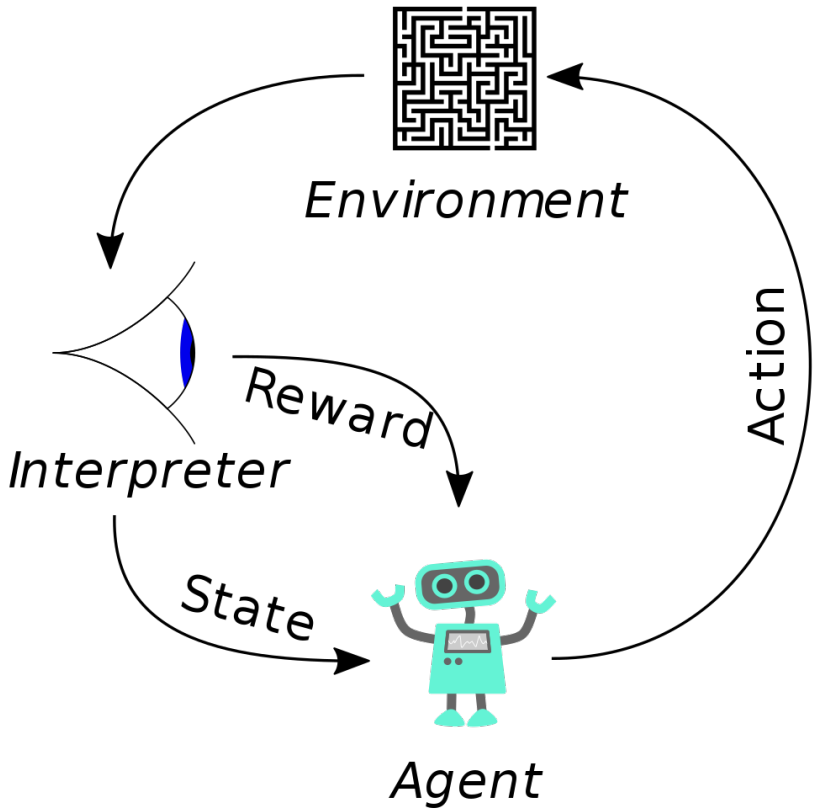


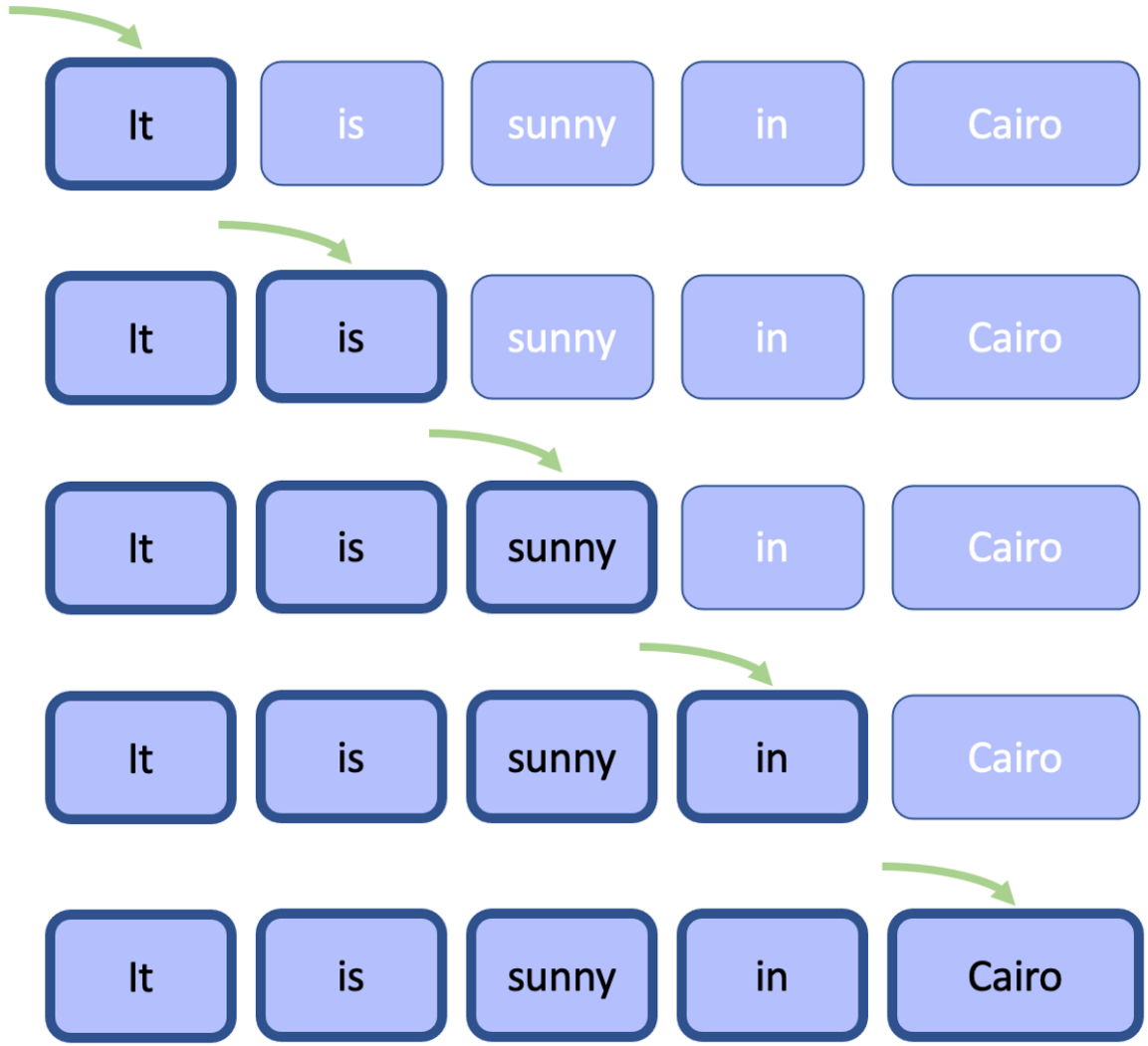
Fig. 2. The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: Ho et al. 2020 with a few additional annotations)

Interacting with Users: Reinforcement Learning



An RL agent interacts with its environment in discrete time steps. At each time t, the agent receives the current state s_t and reward r_t. It then chooses an action a_t, which is subsequently sent to the environment. The environment moves to a new state s_{t+1} and the reward r_{t+1} associated with the transition (s_t, a_t, s_{t+1}) is determined. The goal of an RL agent is to learn a policy π which maximizes the expected cumulative reward. [Source: Wikipedia]

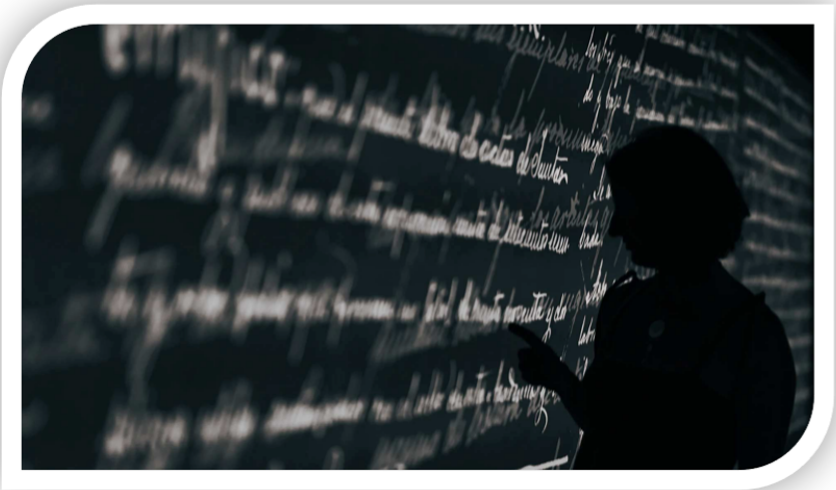
Learning One Token at a Time (Autoregressive Models)



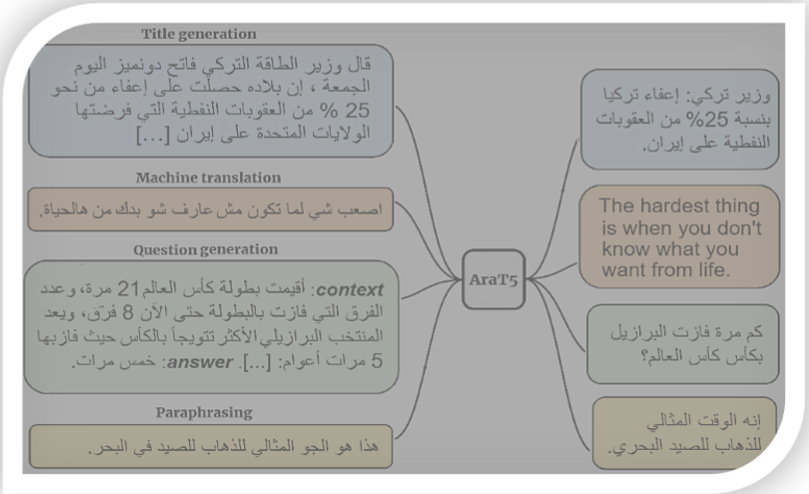
Generative AI

Natural Language Processing

Understanding

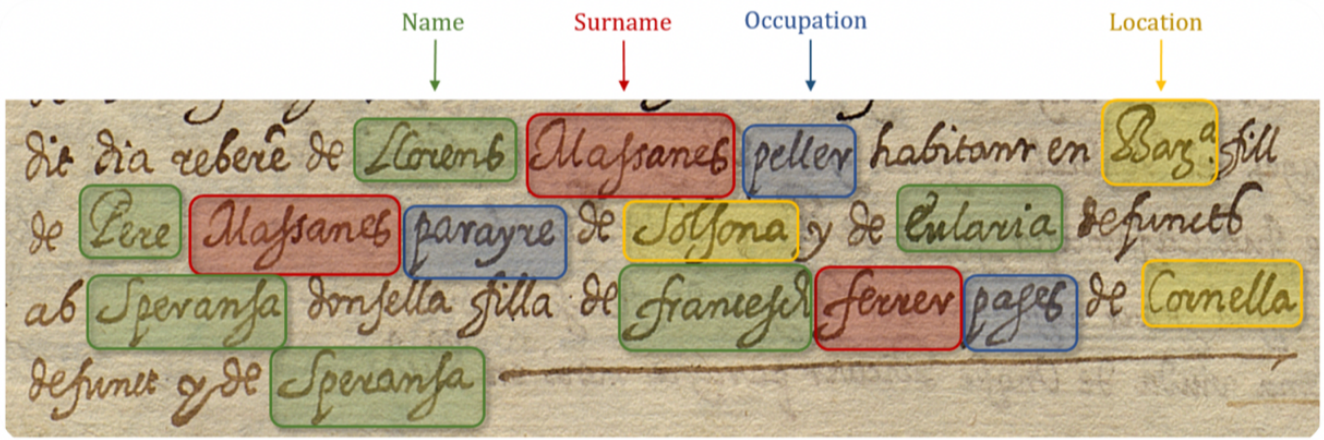


Generation



Example: Named Entity Recognition

Albert Einstein PER Albert Einstein was born in Ulm LOC in Germany LOC on March 14, 1879. Six weeks later the family moved to Munich LOC , where he later on began his schooling at the Luitpold Gymnasium ORG . In 1896 he entered the Swiss Federal Polytechnic School ORG in Zurich LOC to be trained as a teacher in physics and mathematics.



(**Esposalles database**, a marriage license book conserved at the Archives of the Cathedral of Barcelona)
Source: https://rrc.cvc.uab.es.

Named Entity Recognition Tutorials

Named Entity Recognition (NER)

	Category	Descriptions	Link
1	Named Entity Recognition	Introduction to NER and out-of-box solution with Spacy	notebook
2	Named Entity Recognition	Train BiLSTM with PyTorch from Scratch	notebook
3	Named Entity Recognition	Fine-tune BERT with Huggingface	notebook

Figure: [Link]

Emergent Questions

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

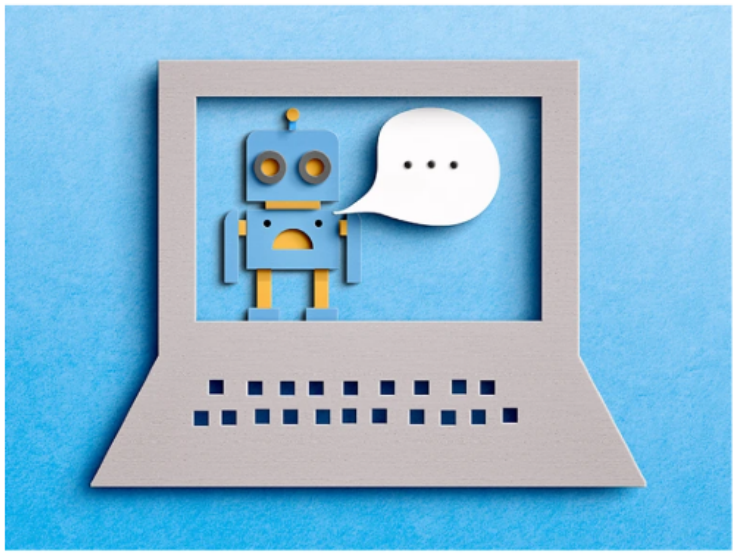
nature › news explainer › article

NEWS EXPLAINER | 13 February 2023

AI chatbots are coming to search engines – can you trust the results?

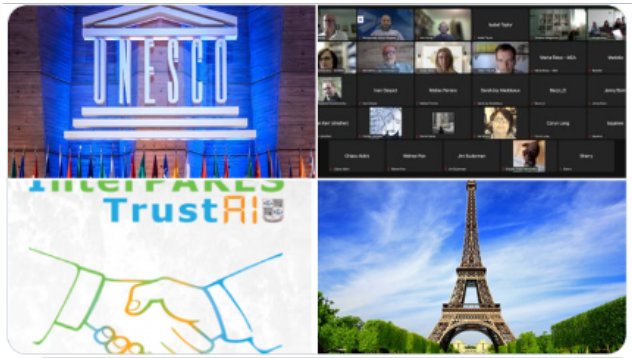
Google, Microsoft and Baidu are using tools similar to ChatGPT to turn Internet search into a conversation. How will this change humanity's relationship with machines?

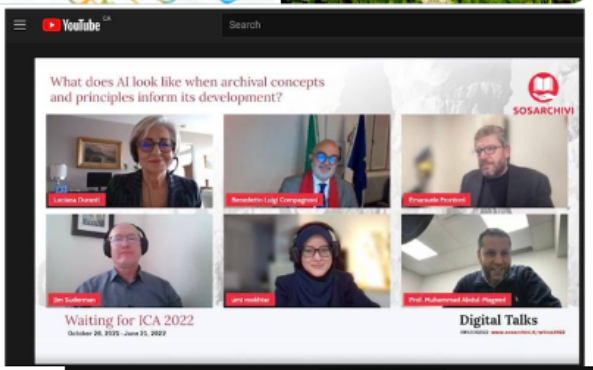
Chris Stokel-Walker



Research has shown that the more human-like a chatbot seems, the more people trust it. Credit: Getty

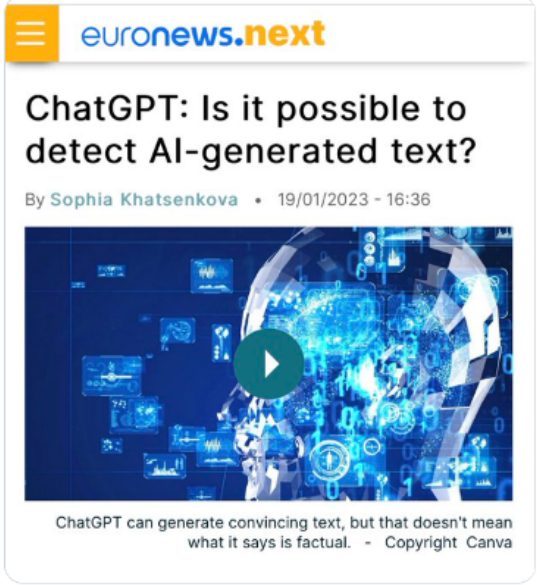
Training & Dissemniation





Media Engagement

technologyreview.com
Why detecting AI-generated text is so difficult (and what to do abou...
Plus: AI models generate copyrighted images and photos of real people.



euronewsde @euronewsde · Feb 18
Wird KI immer besser? Euronews-Journalistin @SKhatsenkova erklärt, wie es kommt, dass der neue Chatbot von Bing einen Reporter auffordert, seine Frau zu verlassen. #TheCube mit @mageed

de.euronews.com
#TheCube deckt auf: Wenn der neue Chatbot von Bing Gefühle zeigt...
Wird KI immer besser? Euronews-Journalistin @SKhatsenkova erklärt, wie es kommt, dass der neue Chatbot von Bing einen Reporter ...

liberation.fr
ChatGPT : comment détecter qu'un texte a été écrit par l'intelligence artificielle ?

Find us

- **Website:** `www.interparestrustai.org`
  Twitter: `www.facebook.com/interparestrust`
  Facebook: `www.facebook.com/interparestrust`