UNIVERSITY OF
LUCERNE

The ABC of Computational Text Analysis

#1 Introduction + Where is the digital revolution?

Alex Flückiger
Faculty of Humanities and Social Sciences
University of Lucerne

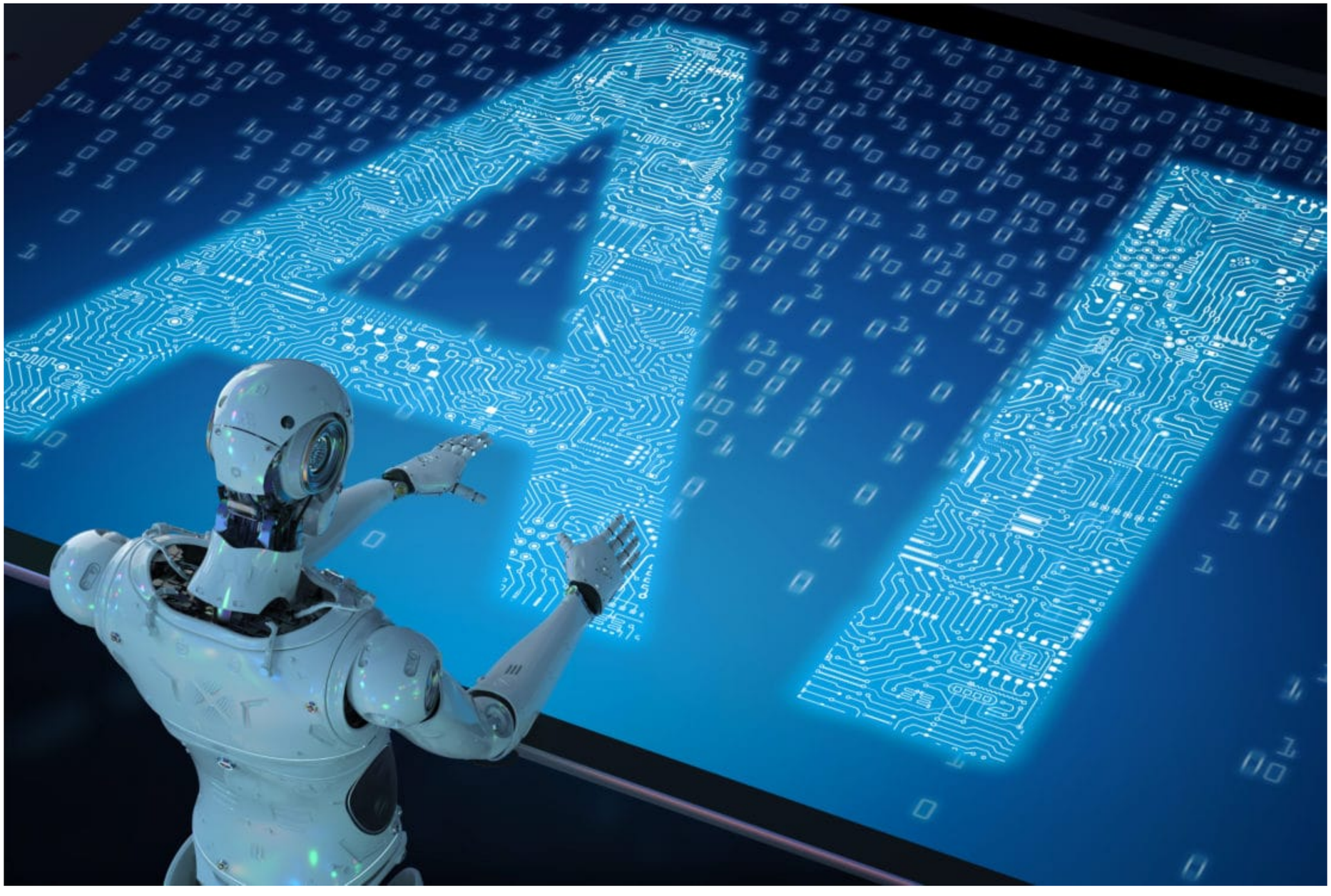
23 February 2023

Outline

1. digital revolution or hype?
2. about us
3. goals of this course

AI: A non-standard Introduction

The World has changed, hasn't it?



An Era of Big Data + AI

Group Discussion

What makes a computer looking intelligent?

AI is a moving target with respect to …

- human capabilities
- technological abilities

Transfer of Human Intelligence

from static machines to more flexible devices

- mimicking intelligent behavior
  - reading + seeing + hearing
  - speaking + writing + drawing
- a sense of contextual perception
- many degrees of freedom

Seeing like a Human?

An image segmentation by Facebook's Detectron2 *(Wu et al. 2019)*

Speaking like a Human?

Speech-to-Text (STT)

Recognizing speech regardless of language, accent, speed, noise etc.

- Check out samples of Whisper (Radford et al. 2022)
- Check demo for Swiss German (Plüss et al. 2021)

Text-to-Speech Synthesis (TTS)

Personalizing voice given an audio sample of 3s

- Check out samples of VALL-E (Wang et al. 2023)

Generative and Multimodal AI

Outsmarting Humans?

ChatGPT is amazing but ...

... it is also a stochastic parrot. 🦜

(Bender et al. 2021)

Can you disenchant ChatGPT?

Experiment with ChatGPT

- What works (surprisingly) well?
- When does it fail?

Generated Images by a Neural Network

https://thisxdoesnotexist.com/

Give me *more!*

Trend towards Multimodality

A storefront with 'Muse' written on it, in front of Matterhorn Zermatt.

A surreal painting of a robot making coffee.

A cake made of macarons in a unicorn shape.

Three dogs celebrating Christmas with some champagne.

Breakthrough by combining language processing and image generation with Muse *(Chang et al. 2023)*

Deepfakes? Yes, they are real!

Editing pictures with Muse using natural language *(Chang et al. 2023)*

Video is just the last barrier...

Synthesize any content with ever increasing quality

- Checkout this demo trailer for authentic dubbing
- Use words and images to synthesize new videos with Gen-1(Esser et al. 2023)

Artificial Intelligence

Subfields

- Natural Language Processing (NLP)
- Computer Vision (CV)
- Robotics

How does Computer Intelligence work?

- interchangeably (?) used concepts

  Artificial Intelligence (AI), Machine Learning (ML), **Deep Learning** (DL)

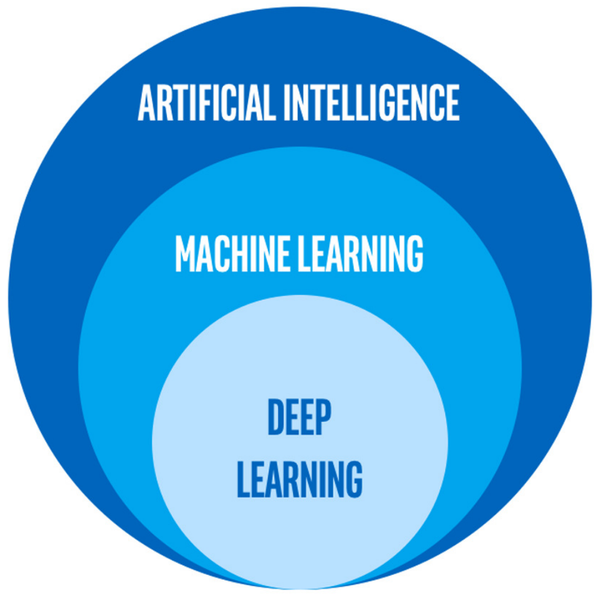
- learn **patterns** from lots of data

  more recycling than genuine intelligence

  theory agnostically

- supervised **training** is the most popular

  learn relation between input and output



AI is also Hype

```
AI = from humankind import solution
```

AI is different to Human Intelligence

Why this matters for Social Science

Computational Social Science

data-driven research

- computational social science (Lazer et al. 2009; Salganik 2017)

  Digital Humanities, Computational History, Data Science

- highly interdisciplinary
- machine learning empowers researchers (Lundberg, Brand, and Jeon 2022)
- early computational history already in 1960s (Graham, Milligan, and Weingart 2015)

Group Discussion

What kind of data is there?

What data is relevant for social science?

- data as traces of social behaviour

  tabular, text, image

- datafication

  sensors of smartphone, digital communication

- much of human knowledge compiled as text

About the Mystery of Coding

coding is like...

- cooking with recipes
- superpowers

Women have coding powers too!

Where the actual Revolution is

Coding is a **superpower** …

- flexible
- reusable
- reproducible
- inspectable
- collaborative

… to tackle complex problems on scale

About us

Personal Example

directed country mentions in UN speeches

Goals of this Course

What you learn

- collect and curate **data**
- **computationally analyze**, interpret, and visualize **texts**

  command line + Python
- **digital literacy** + scholarship
- problem-**solving** capacity

Learnings from previous Courses

- too much content, too little **practice**
- programming can be overwhelming
- **learning** by **doing**, doing by **googling**

Levels of Proficiency

1. **awareness** of today's computational potential
2. **analyzing** existing datasets
3. **creating** + analyzing new datasets
4. applying advanced **machine learning**

How I teach

- computational **practises**
- **critical perspective** on technology
- lecture-style introductions
- hands-on coding sessions
- discussions + experiments in groups

Provisional Schedule

Date	Topic
23 February 2023	Introduction + Where is the digital revolution?
02 March 2023	Text as Data
09 March 2023	Setting up your Development Environment
16 March 2023	Introduction to the Command-line
23 March 2023	Basic NLP with Command-line
30 March 2023 (Zoom)	Learning Regular Expressions
06 April 2023 (Zoom)	Working with (your own) Data
~~13 April 2023~~	*no lecture (Osterpause)*
20 April 2023	Ethics and the Evolution of NLP
27 April 2023	Introduction to Python + VS Code
04 May 2023	Data Analysis of Swiss Media
11 May 2023	NLP with Python
~~18 May 2023~~	*no lecture (Christi Himmelfahrt)*
25 May 2023	NLP with Python II + Working Session
01 June 2023	Mini-Project Presentations + Discussion

🖥️ There are two digital lectures via Zoom.

TL;DR 

You will be tech-savvy…

…yet no programmer applying fancy machine learning

Requirements

- no technical skills required ✅

  self-contained course

- laptop (macOS, Win11, Linux) 

  update system

  free up at least 15GB storage

  backup files

Grading ✍️

- 3 exercises during semester

  no grades (pass/fail)

- **mini-project with presentation**

  backup claims with numbers

  work in teams

  data of your interest

- **optional: writing a seminar paper**

  in cooperation with Prof. Sophie Mützel

Organization

- seminar on Thursday from 2.15pm - 4.00pm

  additionally, streaming via Zoom

- course website KED2023 with slides + information
- readings on OLAT
- communication on OLAT Forum

  forum for everything except personal

  subscribe to notifications

  direct: alex.flueckiger@doz.unilu.ch

Who are you?

Please fill out this questionnaire

Questions?

Reading

Required

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. “Computational Social Science.” *Science* 323(5915):721–23.

(via OLAT)

Optional

Graham, Shawn, Ian Milligan, and Scott Weingart. 2015. *Exploring Big Historical Data: The Historian’s Macroscope*. Open Draft Version. Under contract with Imperial College Press.

online

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Chang, Huiwen, Han Zhang, Jarred Barber, A. J. Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, et al. 2023. “Muse: Text-To-Image Generation via Masked Generative Transformers.” arXiv. https://doi.org/10.48550/arXiv.2301.00704.

Esser, Patrick, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. “Structure and Content-Guided Video Synthesis with Diffusion Models.” arXiv. https://doi.org/10.48550/arXiv.2302.03011.

Graham, Shawn, Ian Milligan, and Scott Weingart. 2015. *Exploring Big Historical Data: The Historian’s Macroscope*. Open Draft Version. Under contract with Imperial College Press. http://themacroscope.org.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. “Computational Social Science.” *Science* 323 (5915): 721–23. https://doi.org/10.1126/science.1167742.

Lundberg, Ian, Jennie E. Brand, and Nanum Jeon. 2022. “Researcher Reasoning Meets Computational Capacity: Machine Learning for Social Science.” *Social Science Research* 108 (November): 102807. https://doi.org/10.1016/j.ssresearch.2022.102807.

Plüss, Michel, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. “Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus.” arXiv. https://doi.org/10.48550/arXiv.2010.02810.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. “Robust Speech Recognition via Large-Scale Weak Supervision.” arXiv. https://doi.org/10.48550/arXiv.2212.04356.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Illustrated edition. Princeton: Princeton University Press. https://www.bitbybitbook.com.

Wang, Chengyi, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, et al. 2023. “Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers.” arXiv.