Synthetic Data, Data Protection and Copyright in an era of Generative AI

Dr. Kalpana Tyagi, Assistant Professor (IP & Competition) &
Founding and Managing Coordinator, The Innovator's Legal Aid Clinic (TILC)
k.tyagi@maastrichtuniversity.nl

Presentation based on research, see here

CIPIL Evening Seminar

24th October 2024

Faculty of Law, University of Cambridge

Title Goes Here

Generative AI

- AI tools that help create content upon prompt
- May be written, audiovisual or even a programming code
- Technical considerations

01

ChatGPT:
- fastest growing digital app
- 3 months of launch, 100 million+ users

02

GenAI: not only extract, but can also contextualize and improvise

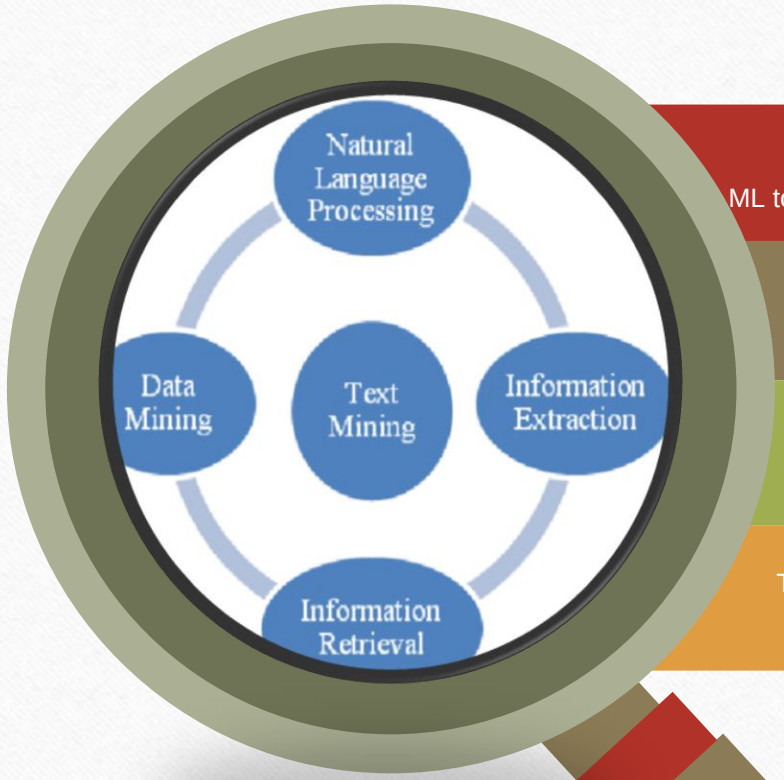
03

Issues on input as well as output side

04

Data – copyright & data protection laws

Unlike earlier versions of NLP, example Word2Vec [that were context independent], recent GenAI models [starting BERT, Bidirectional Encoder Representations from Transformers] are ‘context dependent’



NLP: subset of AI, ML tools that automate natural language functions

Moment of disruption: 2016, when Google transitioned from statistical machine translation to neural machine translation

2017 Google uses ‘transformers’ in NLP-related queries

Transformers disruptive as they transition NLP models from unidirectional to ‘bidirectional’

Context Dependent

BERT & ChatGPT

Designed to pre-train bidirectional representations from unlabeled text by jointly conditioning all sides of the layer and at multiple levels

Pre-GenAI: context independent

Left to right or right to left

But GenAI

Bidirectionality

From Pre-GenAI

To the GENERATIVE AI

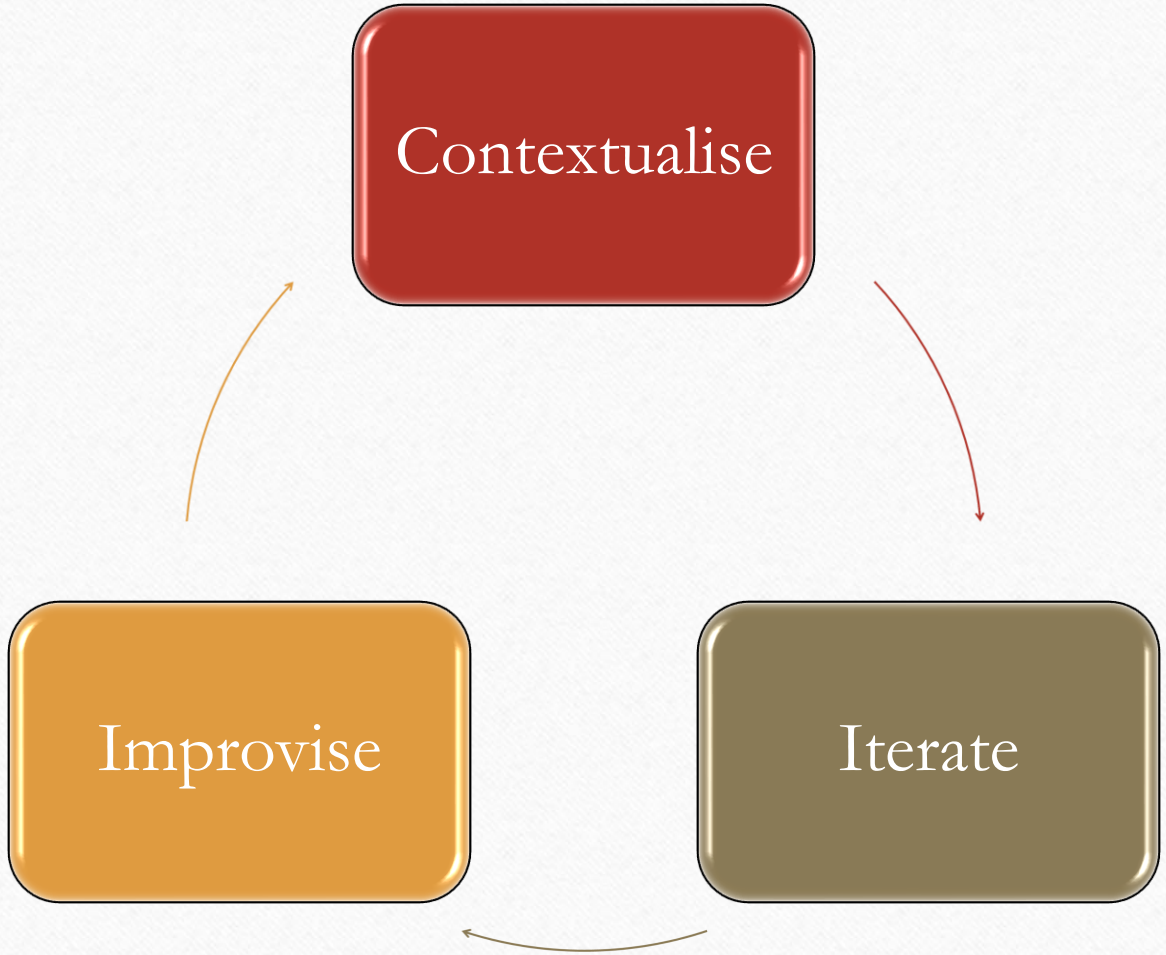
Impact

Abstract, neural network

The CII model

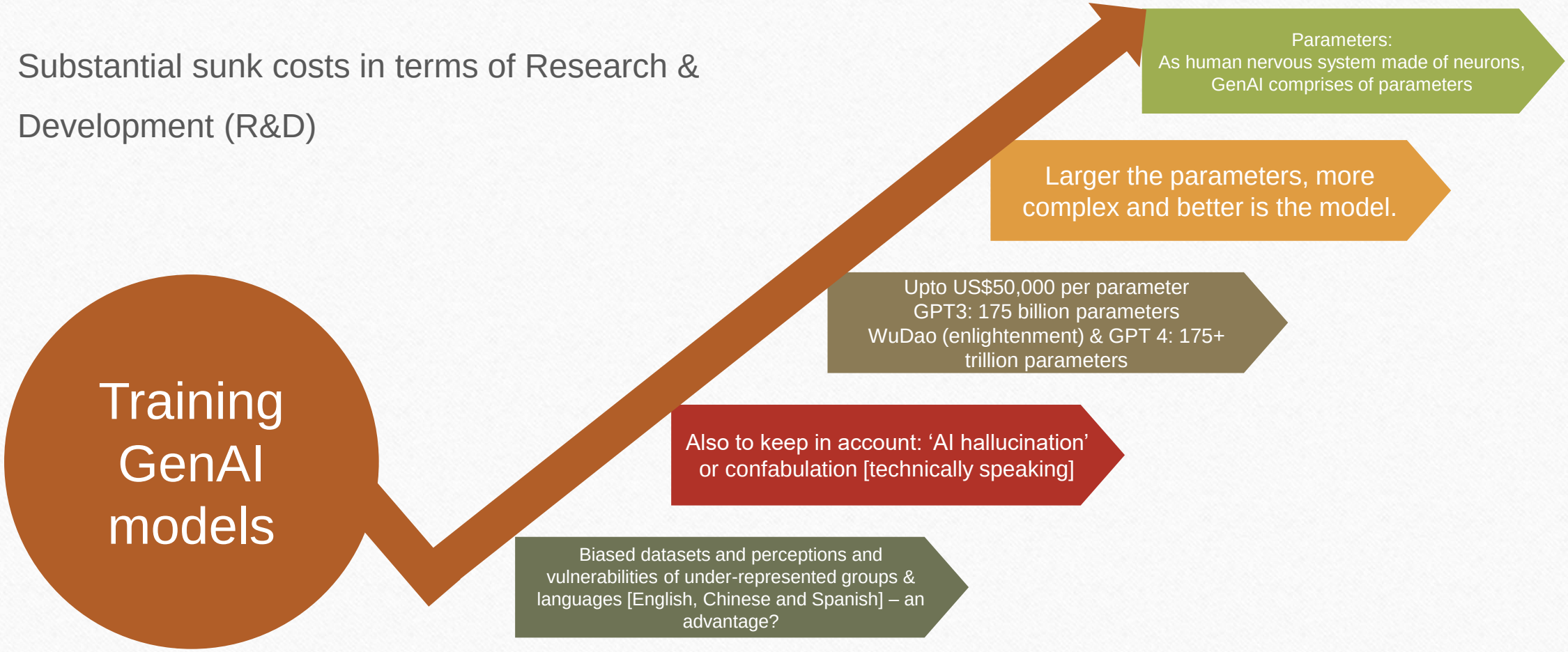
With example:

- *Sentence 1: 'A wise investor makes financially sound investment decisions.'*
- *Sentence 2: 'Contemporary music shows more sound, and less music.'*
- *Sentence 3 (by ChatGPT): 'The symphony of raindrops danced on the rooftop, orchestrating a sound and soothing lullaby for the sleeping city.'*



Law and tech in GenAI

Substantial sunk costs in terms of Research & Development (R&D)



Synthetic Data

- Machine learning algorithms are data hungry
- GenAI and large language models (LLMs) follow the 'neural scaling laws': positive correlation between the model and the size of the training datasets
- Digital gatekeepers, GAMMA – control data, the **training input (data)**, [and other inputs, such as cloud and technical expertise]
- Human-generated data has limited availability
- Current pace of LLM-training, likely exhaustion of 'public human text data between 2026 and 2032' (Villalobos et al)

Synthetic Data

- To overcome the data bottleneck, following technical possibilities:

- Efficient data consumption

- New learning techniques, such as **transfer learning**: Pre-trained parameters used for developing a follow-on model. Example, model trained to identify deep-fake political news can be used to decipher political satire

- **Synthetic data generation**

Synthetic Data

- Artificially-generated data
- Can be visual, written, tabular, audio-visual, graphic
- Key feature: same statistical properties as the original data
- Difference between synthetic and traditional anonymisation techniques: Traditional anonymization – only certain aspects of data (sensitive attributes) anonymised + non-sensitive attributes can be de-anonymised with big data
- Synthetic data that can be reverse-engineered to have the original data **cannot** qualify as a synthetic data.

Synthetic Data

- Different techniques of synthetic data generation:

- SMOTE: Synthetic Minority Over-sampling Technique – to de-bias datasets
- Training IoT systems: need to foresee customer requirements: Amazon Alexa AI team used customer response on the e-commerce platform to create ‘new, similar sentences’
- GenAI enabled IoT, such as Amazon’s DialFRED – mix cycles of human and synthetically-generated data to optimise the system performance and capabilities

Synthetic Data

- Different techniques of synthetic data generation:
- Training GenAI models: ‘development, testing and validation in machine learning systems’
- ChatGPT trained using ‘unsupervised, reinforcement learning coupled with human feedback (RLHF) and semi-supervised learning techniques.

Synthetic Data

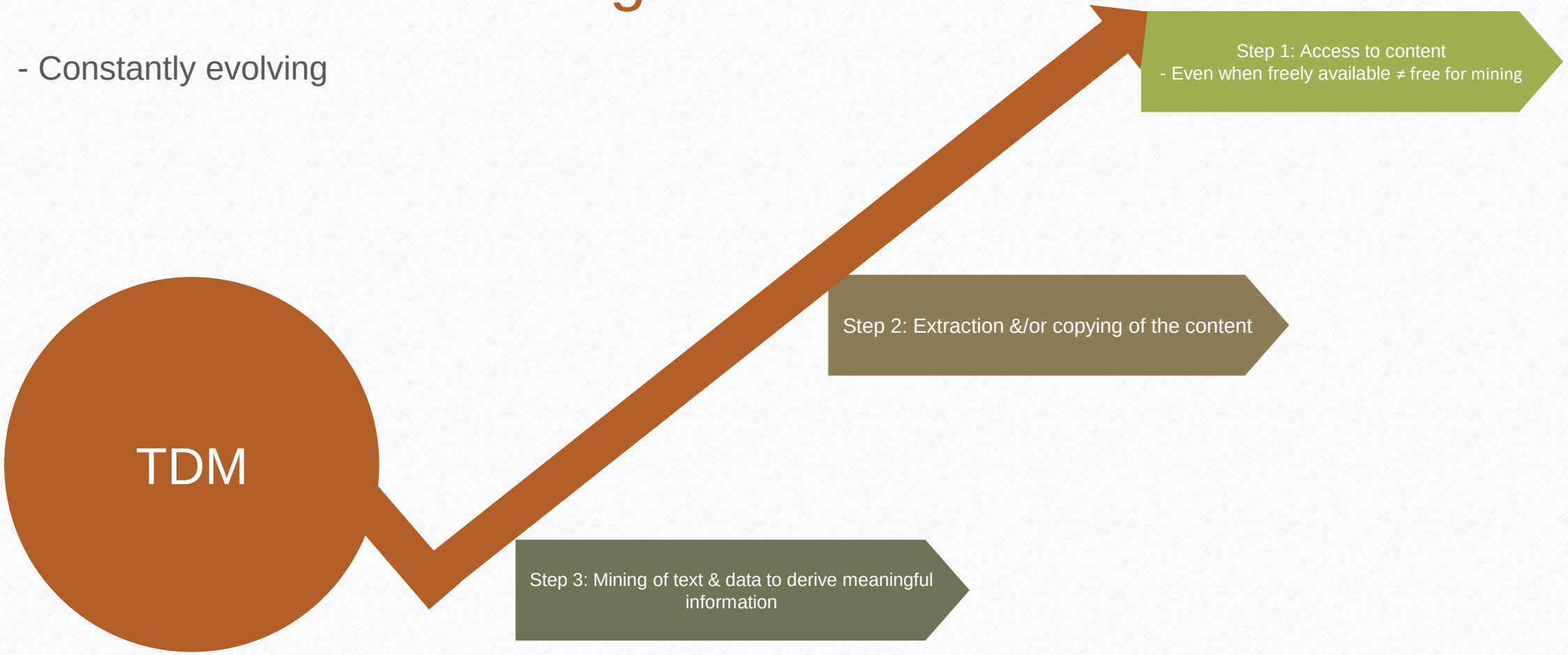
- Fully visible belief networks (FVBN) (Geoffrey Hinton, co-author with Frey and Dayan) – used a **probability-driven approach** to generate synthetic data – slow & limited output generation
- FVBN fine-tuned to create and train advanced models such as WaveNet by DeepMind
- Transition to neural network driven approach + use of transformers – gives 'bidirectionality' to the GenAI models
- GAN and VAE – current and state of the art, synthetic data generation models

Text and Data Mining

- Data is the key
- Data not copyright-protected, ‘creative form’, that is work, that consists of data is copyright-protected + personal data
- TDM: in a constant state of evolution
- Computational linguists, NLP: TDM over 25-30% of the research

Text and Data Mining

- Constantly evolving



TDM in the EU

- Article 3, 2019 CDSM - an exception for the:

  - right of reproduction and extraction under Database Directive & the InfoSoc Directive

  - Press publishers rights under Article 15, 2019 CDSM

- Article 4, 2019 CDSM

  - Also, additionally offers exemption from the right of reproduction and translation, adaptation and alteration of a computer programme under Articles 4(1) (a) and (b), 2009 Computer Programmes Directive [which anyways subject to Article 5(3) – authorize the user, with a lawful access, the possibility to 'observe, study or test the functioning of the programme': Interplay between Article 4, 2019 CDSM (for TDM i.e. TDM dehors research) and Article 5, 2009, CPD (for research purposes)

Robert Kneschke v. LAION

- LAION (Large-scale Artificial Intelligence Open Network), German not-for profit firm, created LAION-5B training dataset
- Dataset: 5.85 billion database filtered image-test pairs – only image descriptions and hyperlinks to the image source; but no images
- Kneschke claims copyright infringement, as scrapping from *bigstock.com* without permission

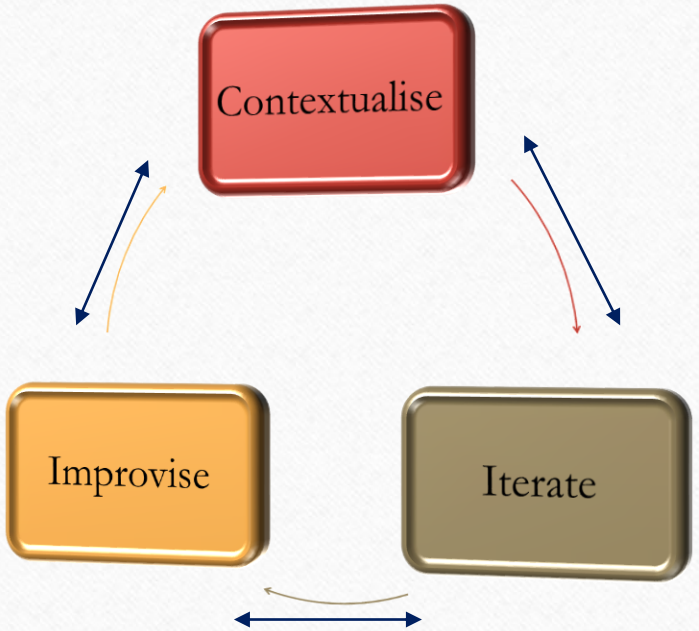
Robert Kneschke v. LAION

- LAIN and Common Crawl's contribution to decoding the black box on which GenAI models are trained
- LAION funding from for-profit firms or output of the TDM, used by commercial for-profit firms – not relevant 'for the assessment under Section 60 UrhG (equivalent Article 4, 2019 CDSM)
- Relevant to court's assessment: Dataset freely available to the public
- Prof. Rosati: dataset made available on the LAION website, an act following after the TDM (under Articles 3 & 4, 2019 CDSM) has already taken place

Robert Kneschke v. LAION

- An obiter dicta (but highly relevant one) on the nature of opt-outs
- 'Digital plain text' as on bigstock – sufficient to communicate opt-out under Article 4(3), 2019 CDSM?
- Article 53(1)(c) 2024 EU AI Act: opt-outs may be exercised in light of the available 'state-of-the-art' technology
- Opt-out in words can be read by NLP.
- Alternatively, develop 'robots.txt' as 'binding' industry standard?
- Why binding nature required now? – Change in market dynamics following the rise of GenAI and requirement of training data.

Synthetic Data, Copyright & TDM



Synthetic Data, Copyright & TDM

- Authors Guild v Open AI
- GenAI tools, ChatGPT here summarises and offers follow-on novels to Grisham's famous work, The King of Torts: 'The Kindgdom of Consequences
- GenAI hallucinates
- Thin line between creativity, hallucination and abstraction, unlike evidence-based facts, where hallucianation may be (relatively) easy to identify

Synthetic Data, Copyright and the EU AI Act

- Article 53(1)(d), 2024 EU AI Act: 'draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.'
- What can possibly comprise this sufficiently detailed summary?
- AI transparency blueprint (by Warso, Gahntz and Keller) – not only synthetically generated data but also the methods used to generate the synthetic output (in addition to other data)

Data Protection and Synthetic Data

- GDPR based in respect for fundamental rights
- Distinction between right to privacy 'old venerable right' & data protection as 'innovation', 'third generation human right' (Prof Erdos)
- GenAI and GDPR: does training involve processing of personal data? & exercise of rights (a challenge) as GenAI models deep learn and black box
- Synthetic data – sufficiently anonymised for GDPR compliance?

Data Protection and Synthetic Data

- Italian DPA (2023) – ChatGPT required a valid legal basis for training personal data
- ChatGPT: Change in policy to 'legitimate interest in "developing, improving or promoting" its services, including the training of its models'
- But what about data of non-users?
- June 2024, EDPS: GenAI model providers may rely on 'legitimate interest…. [especially] with regard to the collection of data' as well as for 'training and validation purposes'.

Data Protection and Synthetic Data

- GenAI models hallucinate, can offer incorrect false information
- Article 4(1)(d), 2016 GDPR – principle of data accuracy
- Inaccuracy in outcome can be averted through accuracy 'throughout the whole lifecycle of the generative AI systems'

Data Protection and Synthetic Data

- Fully synthetic dataset?

– Crucial to ask what specific controls have been put in place (2024 EDPS)

- Data must be evaluated as ‘anonymous, through a proven quality’

Fully synthetic

Can qualify as pseudonymous or anonymised data under Article 4(5) GDPR

Data Protection and Synthetic Data

- Partially synthetic dataset?

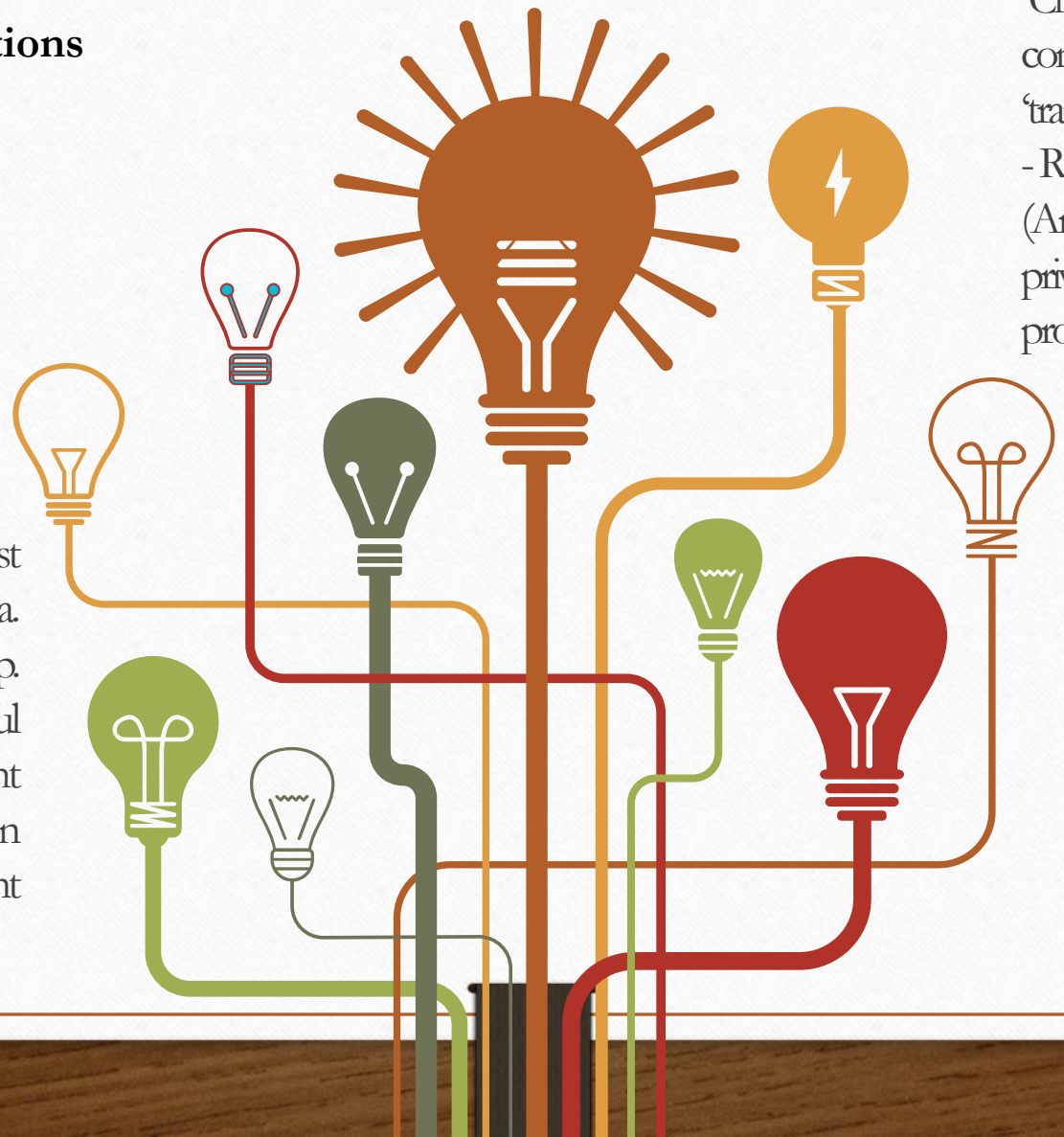
– Partly synthetic, or combination of synthetic and human generated

- Subject to GDPR, as this may involve processing of personal data (Article 4(1), GDPR)
- Thus, required compliance with Article 5, GDPR – namely, lawfulness, fairness, transparency, purpose limitation, data minisation, storage limitation, accountability, accuracy, integrity and confidentiality.

Summary & suggestions

Innovation driven synthetic data as enabler of different rights and competing interests at stake

Charter of Fundamental Rights compliant framework, 'transparency' is the key
- Right to author remuneration (Article 17(2), CFR) + right to privacy (Article 7, CFR) and data protection (Article 8, CFR)



Copyright and data protection close interplay in GenAI

Synthetic Data to co-exist alongside human-generated data. Provisions under the AI Act, esp. 53(1)(c) and (d) can be useful complements effective copyright and data protection enforcement