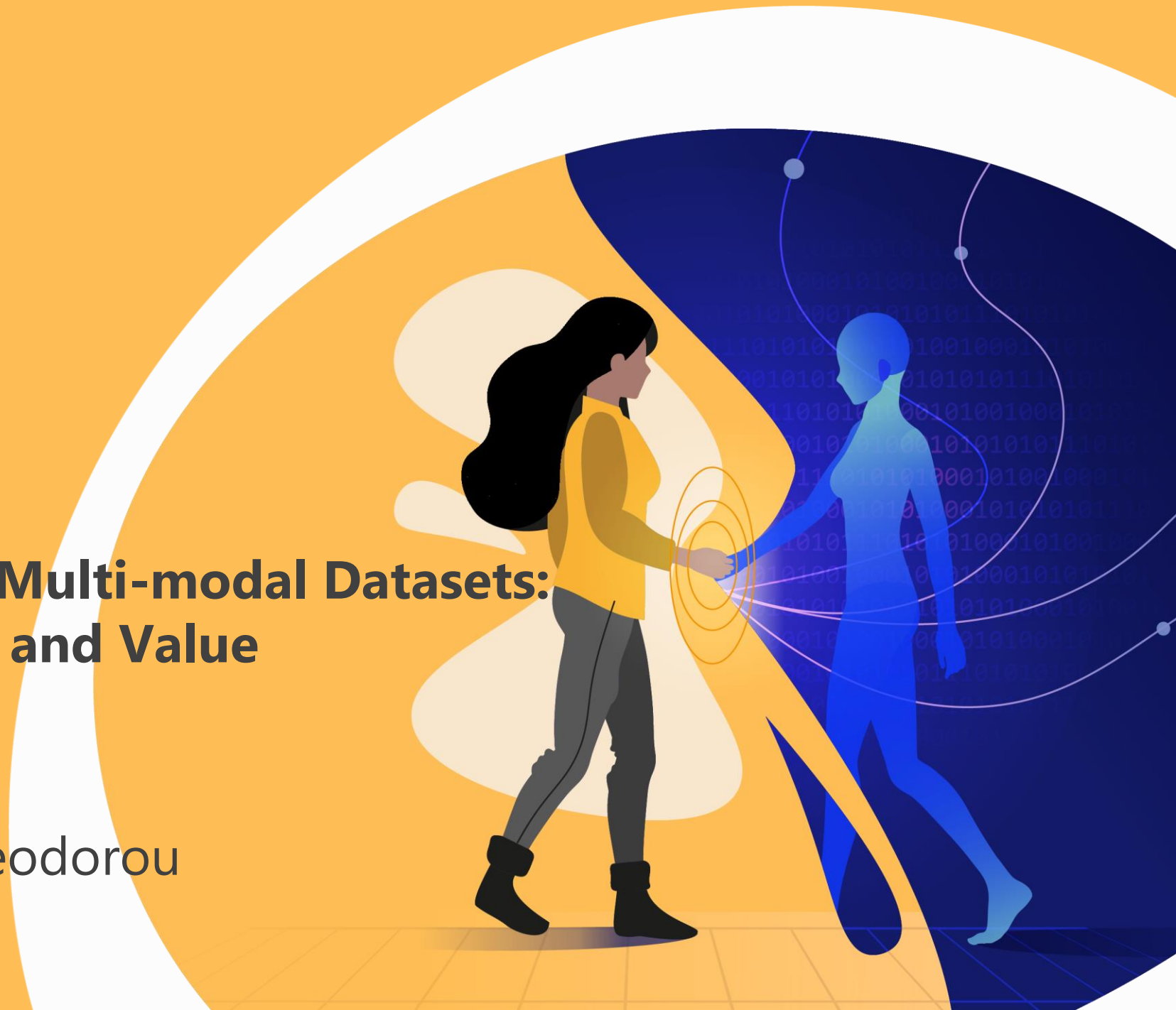




Protecting Privacy in Multi-modal Datasets: Balancing Anonymity and Value

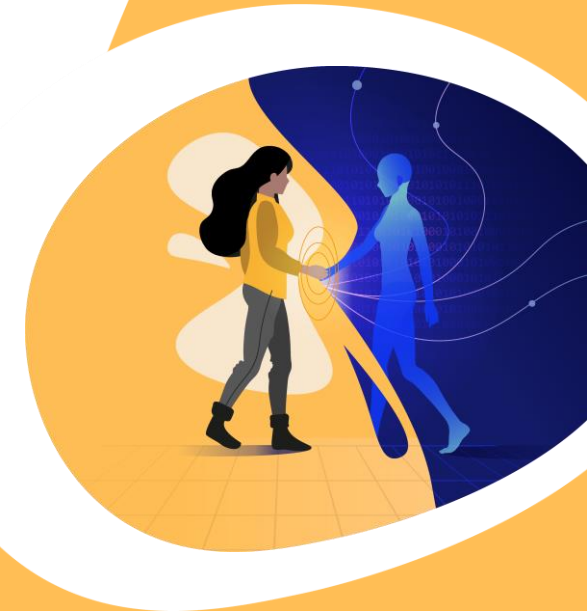
Konstantinos Chatzitheodorou
CTO, Pangeanic



Pangeanic

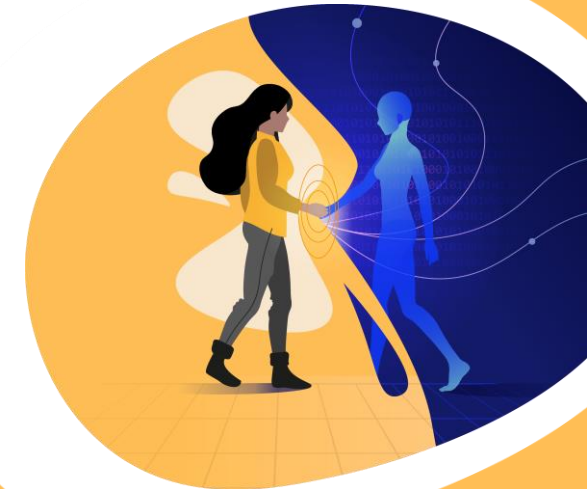
- Deep Adaptive MT
- Text anonymization
- Text classification
- Text summarization
- Sentiment analysis
- Named-entity recognition
- Language detection

- Data for AI
 - Monolingual, multilingual, multimodal



Konstantinos Chatzitheodorou

- MSc in Informatics for humanities
- PhD in Computational Linguistics
- PostDoc in Quality of Translation using AI

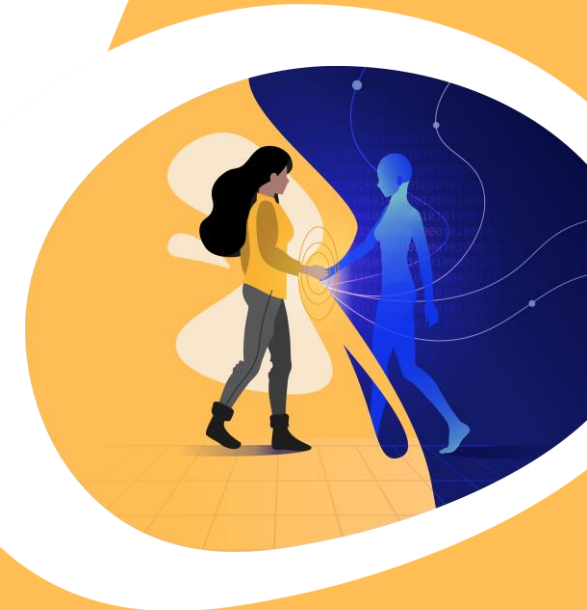
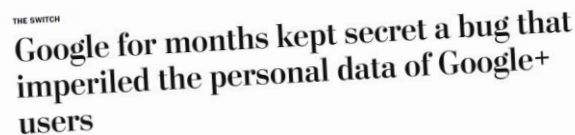


www.pangeanic.com



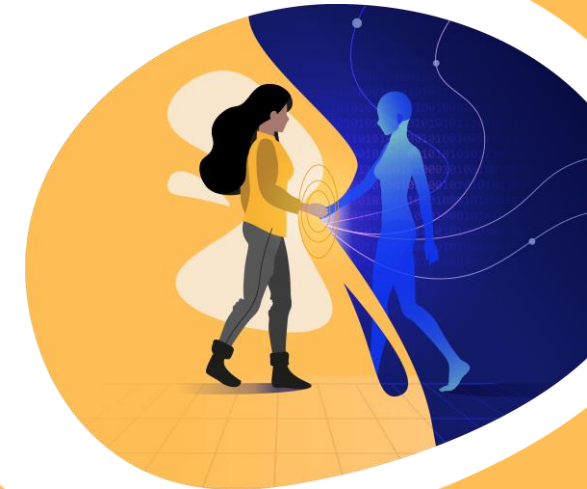
Importance of Privacy protection

- Preserves individuals' personal information, autonomy, and anonymity.
- Builds trust between organizations and individuals.
- Mitigates privacy risks and unauthorized access.
- Demonstrates legal and ethical responsibilities.



Agenda

- Privacy Risks in Multi-modal Datasets
- Ethical Considerations
- Anonymization Techniques
- Strengths, Limitations, and Trade-offs
- Role of Regulations and Policies
- Future Directions
- Conclusion & Q&A

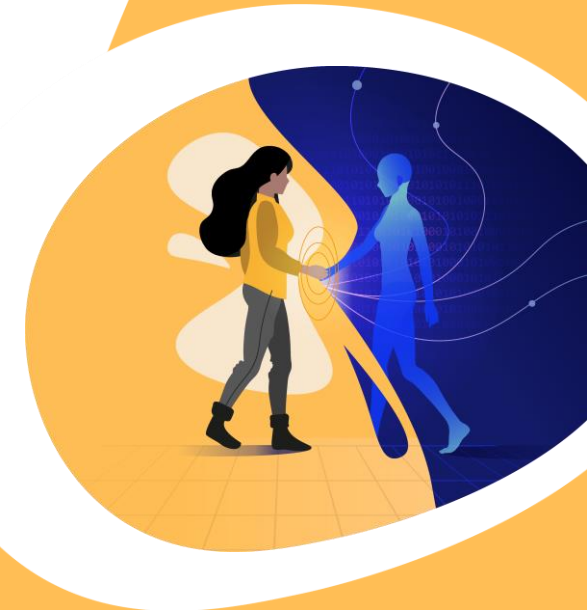




Privacy Risks in Multi-modal Datasets

Multi-modal Datasets

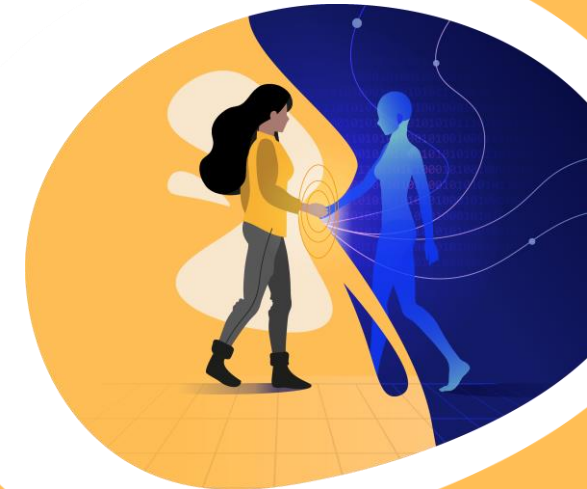
- Multimodal datasets combine different types of data, such as text, audio, image, and video.
- They offer a more comprehensive understanding by capturing multiple aspects of the same subject or event.
- Each modality provides unique information, allowing for a richer representation of real-world phenomena.
- Multimodal datasets find applications in areas like natural language processing, computer vision, audio analysis, etc.



Risks of Multi-modal Datasets and Identification

- Risks of re-identification and unintended disclosures.
- Possibility of combining modalities for identification.
- Implications for privacy and potential harm to individuals.

??? bought the social networking site where users broadcast short posts in October 2022 for \$43.2 billion.



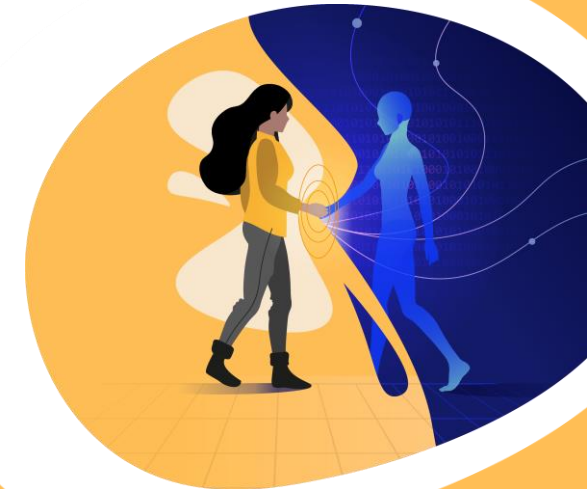
Ethical Considerations and Consequences

- Respecting individuals' rights and autonomy.
- Consequences of mishandling sensitive data.
- Breaches of trust and reputational damage.
- Legal ramifications and potential discrimination.



Role of Regulations and Policies

- **GDPR (EU)**
Emphasizes the rights of individuals, requires consent for data processing, and mandates the implementation of privacy measures such as data anonymization and pseudonymization.
- **CCPA (California)**
Requires businesses to disclose data collection practices and allows consumers to opt-out of the sale of their personal data.
- **LGPD (Brazil)**
Aims to protect privacy and foster responsible data handling practices.
- **APPI (Japan)**
Establishes guidelines for data protection, including consent requirements and measures to ensure the appropriate anonymization and de-identification of personal data.





Anonymization Techniques

Anonymization Techniques - Text Data

- Redaction: Removing or masking sensitive information.
- Generalization: Aggregating or replacing specific values with broader categories.
- Perturbation: Introducing noise or randomization to modify the data.
- Tokenization: Replacing identifiable information with anonymous tokens.
- Differential Privacy: Adding controlled noise to preserve privacy.



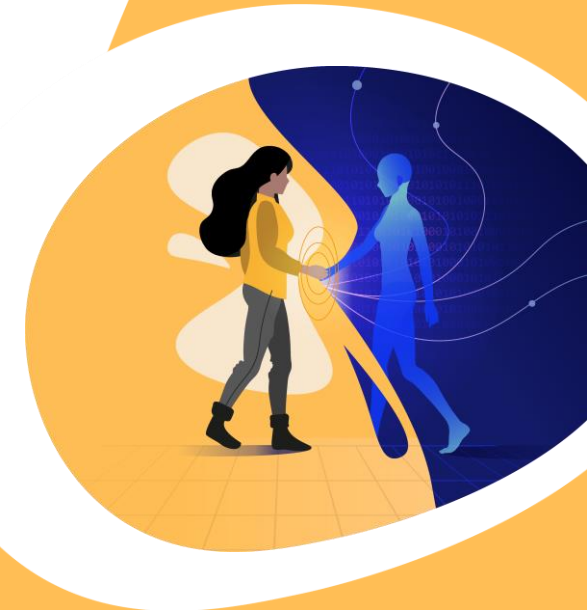
Anonymization Techniques - Text Data

- Redaction
 - **Elon Musk**'s credit card number is **3742-4545-5400-126**.
 - **John Smith**'s credit card number is **1234-5678-9012-456**.
- Generalization
 - The age distribution of the survey participants is as follows: **25, 28, 32, 40, 45, and 50**.
 - The age distribution of the survey participants is as follows: **under 30, 30s, and 40s**.
- Perturbation
 - The user's location data: Latitude: **40.7128**, Longitude: **-74.0060**.
 - The user's location data: Latitude: **40.7132**, Longitude: **-74.0065**.
 - Tokenization
 - His email address is **k.chatzitheodorou@pangeanic.com**.
 - His email address is **[EMAIL]**.
- Differential Privacy
 - Out of 100 respondents, **75** have voted in favor of the proposal.
 - Out of 100 respondents, approx. **73-77** have voted in favor of the proposal.

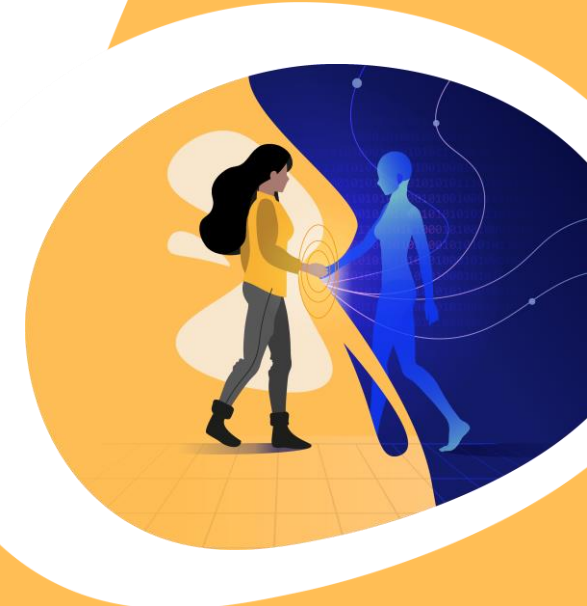
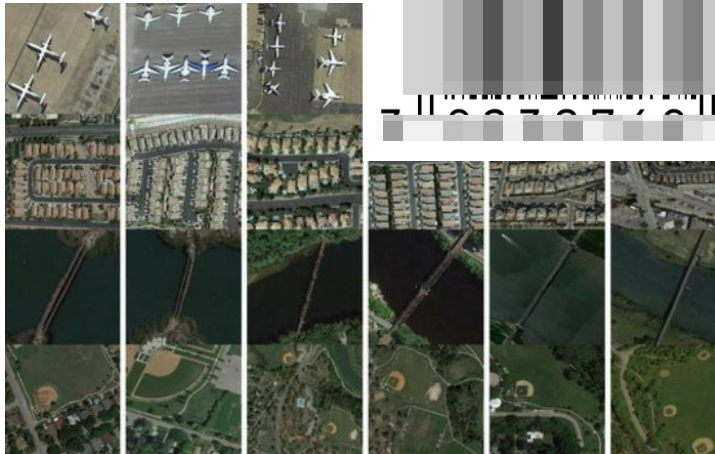
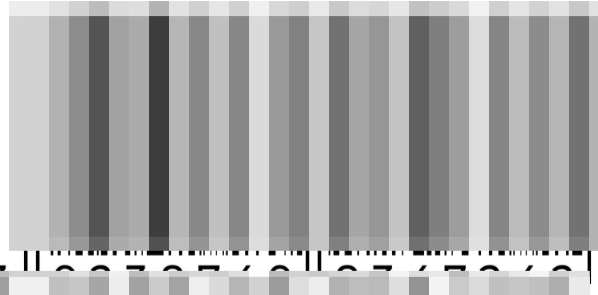
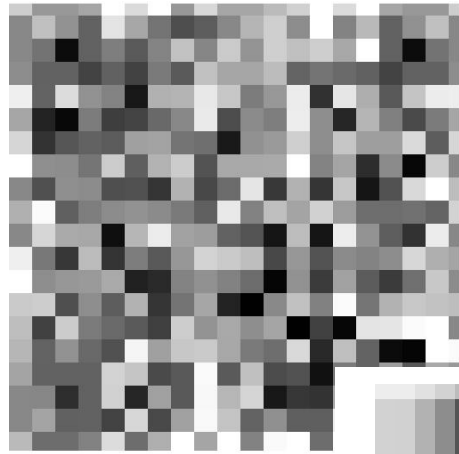


Anonymization Techniques - Image Data

- Image blurring: Blurring or pixelating sensitive regions.
- Image cropping: Removing or cropping out identifiable elements.
- Object removal: Erasing or obscuring specific objects or individuals.
- Metadata removal: Stripping metadata that may reveal sensitive information.
- Image aggregation: Combining multiple images to protect individual identities.

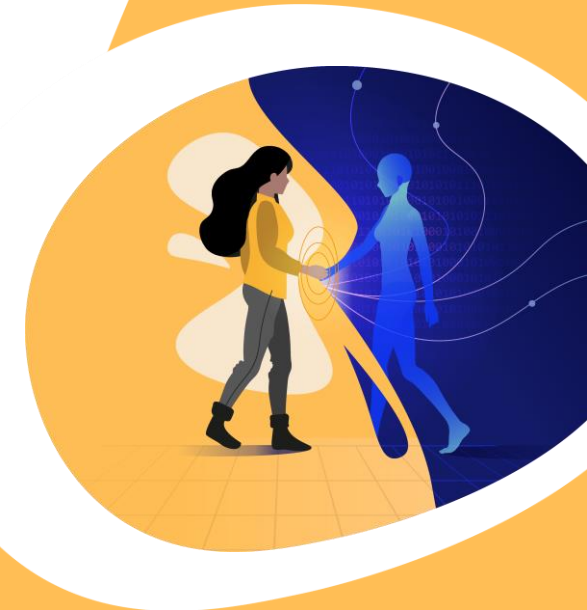


Anonymization Techniques - Image Data



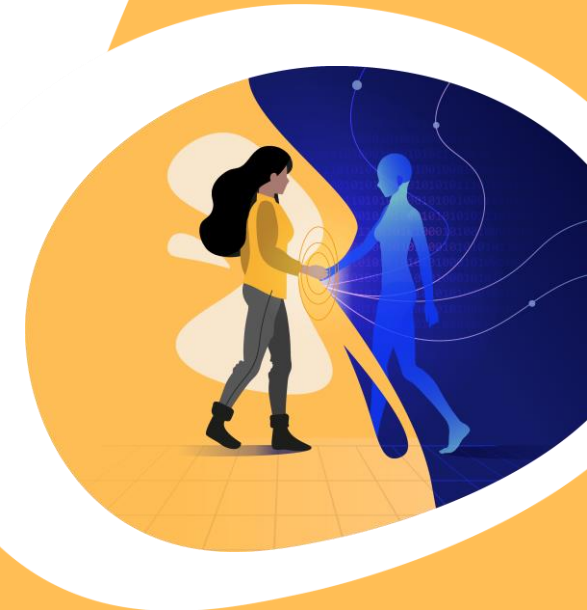
Anonymization Techniques - Audio Data

- Audio redaction: Removing or obfuscating sensitive content.
- Audio encryption: Protecting the audio data with encryption algorithms.
- Audio distortion: Introducing noise or altering the audio characteristics.
- Voice transformation: Modifying the speaker's voice for anonymity.
- Audio aggregation: Combining multiple audio sources for privacy.



Anonymization Techniques - Video Data

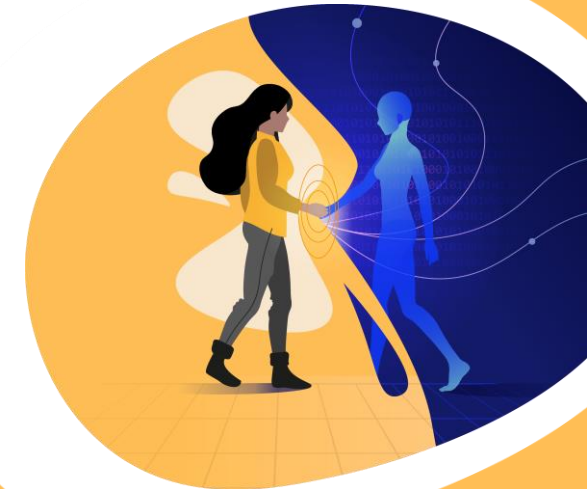
- Face blurring: Blurring or pixelating faces to protect identities.
- Object tracking: Removing or obscuring specific objects or individuals throughout the video.
- Scene substitution: Replacing sensitive scenes or backgrounds to maintain privacy.
- Time-lapse: Speeding up or slowing down the video to mask identities.
- Video aggregation: Combining multiple videos to protect individual identities.



Anonymization Techniques – Video/Audio Data



 pangeanic



www.pangeanic.com

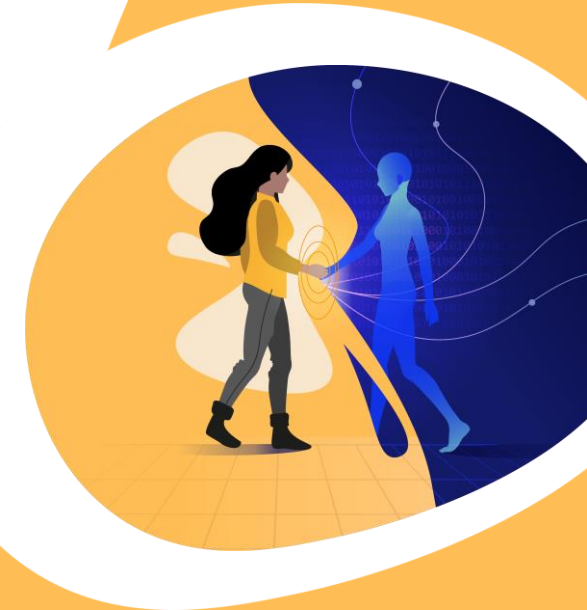




Strengths, Limitations, and Trade-offs

Strengths

- Anonymization techniques help preserve privacy by obscuring or removing sensitive information from datasets.
- They allow organizations to comply with privacy regulations (e.g., GDPR, CCPA, LGPD, APPI) and ethical guidelines when working with personal data.
- Anonymization techniques enable the utilization of valuable datasets for research, analysis, and innovation while respecting individuals' privacy.



Limitations

- Anonymization techniques may not guarantee 100% anonymity, especially with the advancement of re-identification methods and AI algorithms.
- Aggressive anonymization can sometimes compromise the utility and usefulness of the data, making it less valuable for analysis and research.
- There can be challenges in balancing the level of anonymization required to protect privacy with maintaining the necessary information for meaningful insights.



Trade-offs

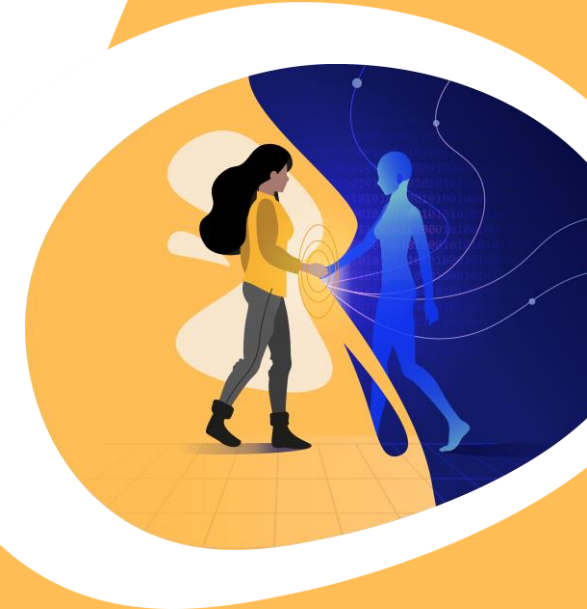
- The choice of anonymization technique depends on the specific data type and the desired level of privacy protection.
- Some techniques, such as aggressive data redaction or distortion, may offer higher privacy but at the cost of reduced data utility and analytical value.
- Striking the right balance between privacy and utility requires careful consideration of the data's context, sensitivity, and intended use.





Pangeanic's Masker

At Pangeanic we simplify Data Anonymization



* Process: Anonymization

* Source: English

Target: English

Team/Engine: Masker

Glossaries:

Profiles:

* Type: Blacklining

Tags:

In the recent company meeting, it was revealed that Emily Thompson from the marketing department will be promoted to the position of Senior Marketing Manager. This decision was based on her outstanding performance, dedication, and successful execution of various marketing campaigns. The announcement was met with applause and congratulations from colleagues, recognizing Emily's contributions to the company's growth. As part of her new role, Emily will be responsible for overseeing strategic marketing initiatives and collaborating with the sales team to drive revenue. The entire team is excited to work under Emily's leadership and is looking forward to achieving new milestones together.

In the recent company meeting, it was revealed that ██████████ from the marketing department will be promoted to the position of ██████████ Marketing ██████████. This decision was based on her outstanding performance, dedication, and successful execution of various marketing campaigns. The announcement was met with applause and congratulations from colleagues, recognizing ██████████'s contributions to the company's growth. As part of her new role, ██████████ will be responsible for overseeing strategic marketing initiatives and collaborating with the sales team to drive revenue. The entire team is excited to work under ██████████'s leadership and is looking forward to achieving new milestones together.



Process

www.pangeanic.com



Pangeanic Solves ChatGPT's Privacy Problem With its Anonymization Layer, Masker



GPT

rephrase the paragraph "In the recent company meeting, it was revealed that Emily Thompson from the marketing department will be promoted to the position of Senior Marketing Manager. This decision was based on her outstanding performance, dedication, and successful execution of various marketing campaigns. The announcement was met with applause and congratulations from colleagues, recognizing Emily's contributions to the company's growth. As part of her new role, Emily will be responsible for overseeing strategic marketing initiatives and collaborating with the sales team to drive revenue. The entire team is excited to work under Emily's leadership and is looking forward to achieving new milestones together."

During the recent meeting held by the company, it was disclosed that a member of the marketing department, identified as [[PER:1]], will be taking on the position of [[PER:2]] Marketing [[PROF:1]]. This decision was made based on her exceptional performance, unwavering dedication, and successful execution of a variety of marketing campaigns. The news was met with enthusiastic applause and congratulations from co-workers, as they acknowledged [[PER:3]]'s contributions in the company's growth. In her new role, [[PER:3]] will be in charge of supervising key marketing initiatives and collaborating with the sales team to boost revenue. The entire team is excited to have [[PER:3]] as their leader, and they are eager to work together to achieve new milestones.

Enter your message



www.pangeanic.com



VALENCIA

MADRID

LONDON

BOSTON

NEW YORK

HONG KONG

TOKYO



Thank you!

www.pangeanic.com

