CLAIMS

What is claimed is:

1. A computer-based system for personal data classification comprising:

a plurality of records stored in one or more data repositories in a computer network or cloud infrastructure, wherein the plurality of records comprise structured, semi-structured and unstructured records; and

a multi-module processing system comprising:

an entity extraction module configured to extract personal data from the plurality of records in the one or more data repositories in the computer network or cloud infrastructure, wherein the entity extraction module is configured to transform text in the plurality of records into numerical vectors using word embedding techniques and provide the numerical vectors to deep learning models to extract the personal data from the plurality of records in the one or more data repositories in the computer network or cloud infrastructure;

a linkage module coupled to the entity extraction module and configured to use graph-based methodology to link the personal data to one or more individuals, wherein the group-based methodology comprises building a knowledge base and using graph mining on a knowledge graph in the knowledge base; and

a purpose prediction module coupled to the linkage module and comprising:

a feature extraction module, wherein the feature extraction module is configured to extract both context features and record features from the plurality of records in the one or more data repositories in the computer network or cloud infrastructure, wherein the record features comprise an extension of the record representing a document type of the record and a language of the record; and

a purpose of processing prediction module configured to predict a unique purpose or multiple purposes of processing of the personal data in response to the context features and the record features,

wherein the purpose prediction module is further configured to use the link of the personal data to the one or more individuals to predict the unique purpose or the multiple purposes of processing of the personal data.

2. The system for personal data classification in accordance with Claim 1 further comprising an anomaly detector configured to detect data breaches in response to the predicted purpose of processing the personal data.

3. The system for personal data classification in accordance with Claim 1 wherein the entity extraction module is configured to use unsupervised deep learning to extract the personal data from the plurality of records, and wherein the deep learning models to which the numerical vectors are provided by the entity extraction module comprise deterministic learning methodology.

4. The system for personal data classification in accordance with Claim 1 wherein the entity extraction module is configured to perform entity recognition to extract more than fifty entity types with their further characterization.

5. The system for personal data classification in accordance with Claim 1 wherein the record features further include user rights, document date and document owner.

6. A computer-implemented method for auto-labelling of personal data comprising:

scanning one or more documents stored in one or more data repositories in a computer network or cloud infrastructure, the one or more documents comprising structured, semi-structured or unstructured documents;

performing entity recognition in the structured, semi-structured and unstructured documents by extracting personal data from the one or more documents, the personal data extracted from the one or more documents by transforming text in the one or more documents into numerical vectors using word embedding techniques and providing the numerical vectors to deep learning models to extract the personal data;

extracting features from the structured, semi-structured and unstructured documents, wherein the features extracted from the structured, semi-structured and unstructured documents comprise metadata extensions and language of the one or more documents; and

personal data processing prediction comprising unsupervised auto-labelling of personal data from documents in one or more data repositories in a computer network or cloud infrastructure, wherein the unsupervised auto-labelling reuses a text summarization methodology and includes key-phrase aggregation and linking techniques to predict a purpose of processing topic for the personal data.

7. The computer-implemented method in accordance with Claim 6 wherein the extracting features comprises extracting more than fifty entity types and further categorization of personal data in the one or more documents.

8. The computer-implemented method in accordance with Claim 6 further comprising:

supervised and unsupervised machine learning model training for identifying relationships between personal data entities in the extracted data; and

calculating a probability of a linking of the personal data entities based on the identified relationships for the unsupervised auto-labelling of the personal data.

9. The system for personal data classification in accordance with Claim 1 further comprising an interactive user interface coupled to the linkage module, wherein the linkage module is further configured to generate one or more visualizations of linkage of the extracted personal data and the one or more individuals, and wherein the interactive user interface is configured to present one or more visualizations of the linkage of the extracted personal data and the one or more individuals for other machine learning tasks including document clustering and classification.