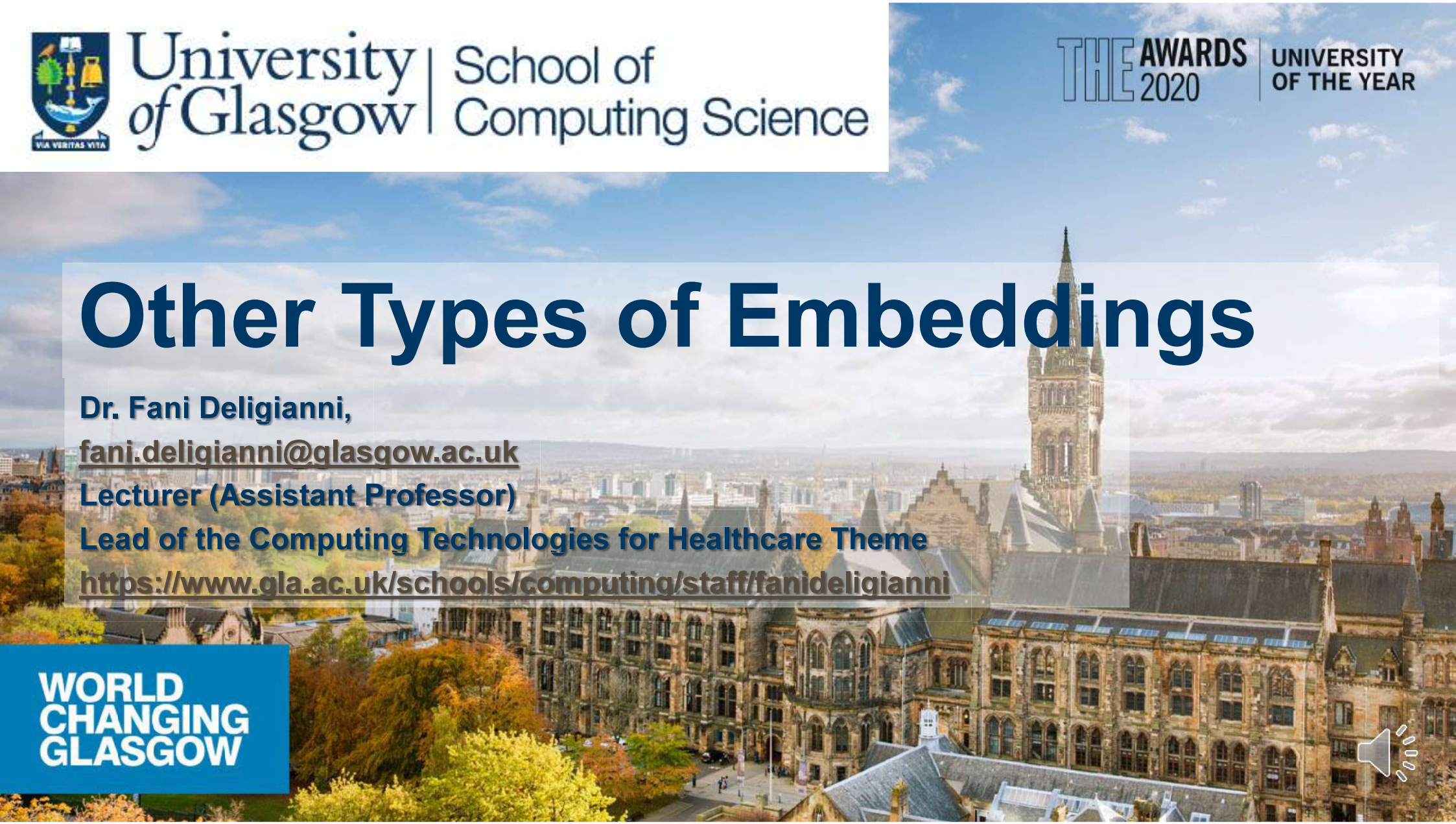

Other Types of Embeddings

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

Dirty data problem

- Non-standardised categorical variables
- Frequency of different categories differ by several orders of magnitude
- Examples of ‘dirty’ data:
  - Typographical errors
  - Extra data recorded non consistently
  - Abbreviations/Aliases
  - Encoding format
  - Special characters
  - Concatenated hierarchical data
- Require knowledge-engineering

Approaches in Database Cleaning

- Database queries with inexact matching
- Remove duplicates
- Database linkage
- Fuzzy matching

Similarity between strings

- Levenshtein-ratio

$$sim_{lev-ratio}(s_1, s_2) = 1 - \frac{d_{lev}(s_1, s_2)}{|s_1| + |s_2|}$$

- Jaro-Winkler

$$d_{jaro}(s_1, s_2) = \frac{m}{3|s_1|} + \frac{m}{3|s_2|} + \frac{m - t}{3m}$$

- N-gram

$$sim_{n-gram}(s_1, s_2) = \frac{|n - grams(s_1) \cap ngrams(s_2)|}{|n - grams(s_1) \cup ngrams(s_2)|}$$

Similarity Encoding

- Generalised one-hot encoding to account for similarities in the categories of a categorical variable
- Replace the categorical variable with a vector

$$x^i = [sim(d^i, d_1), sim(d^i, d_2), \ldots, sim(d^i, d_k)]$$

Dimensionality Reduction: Random Projection

$$x^i = [sim(d^i, d_1), sim(d^i, d_2), \ldots, sim(d^i, d_k)]$$

- Reduced representation that approximates well distances between vectors
- It requires estimating similarity between all categories
  - Random Fourier Features
  - Random Binning Features

Dimensionality Reduction - Clustering

$$x^i = [sim(d^i, d_1), sim(d^i, d_2), \ldots, sim(d^i, d_k)]$$

- The most frequent categories are more dominant
- Reduced representation via clustering ie. K-means
- Clustering can choose elements in the category set
- Choose those elements that are closest to the centre of the cluster

Summary

- One-hot-encoding treats all categories of categorical variables as independent and equally distanced
- Similarity encodings represent the similarity between subcategories
- Data are not required to be clean before similarity encoding

References

- Cerda et al. ‘Similarity encoding for learning with dirty categorical Variables’, Special Issue of the ECML PKDD, 2018.
- Chen et al. ‘Representation Learning for Electronic Health Records: A Survey’, J. Phys Conf, 2020.