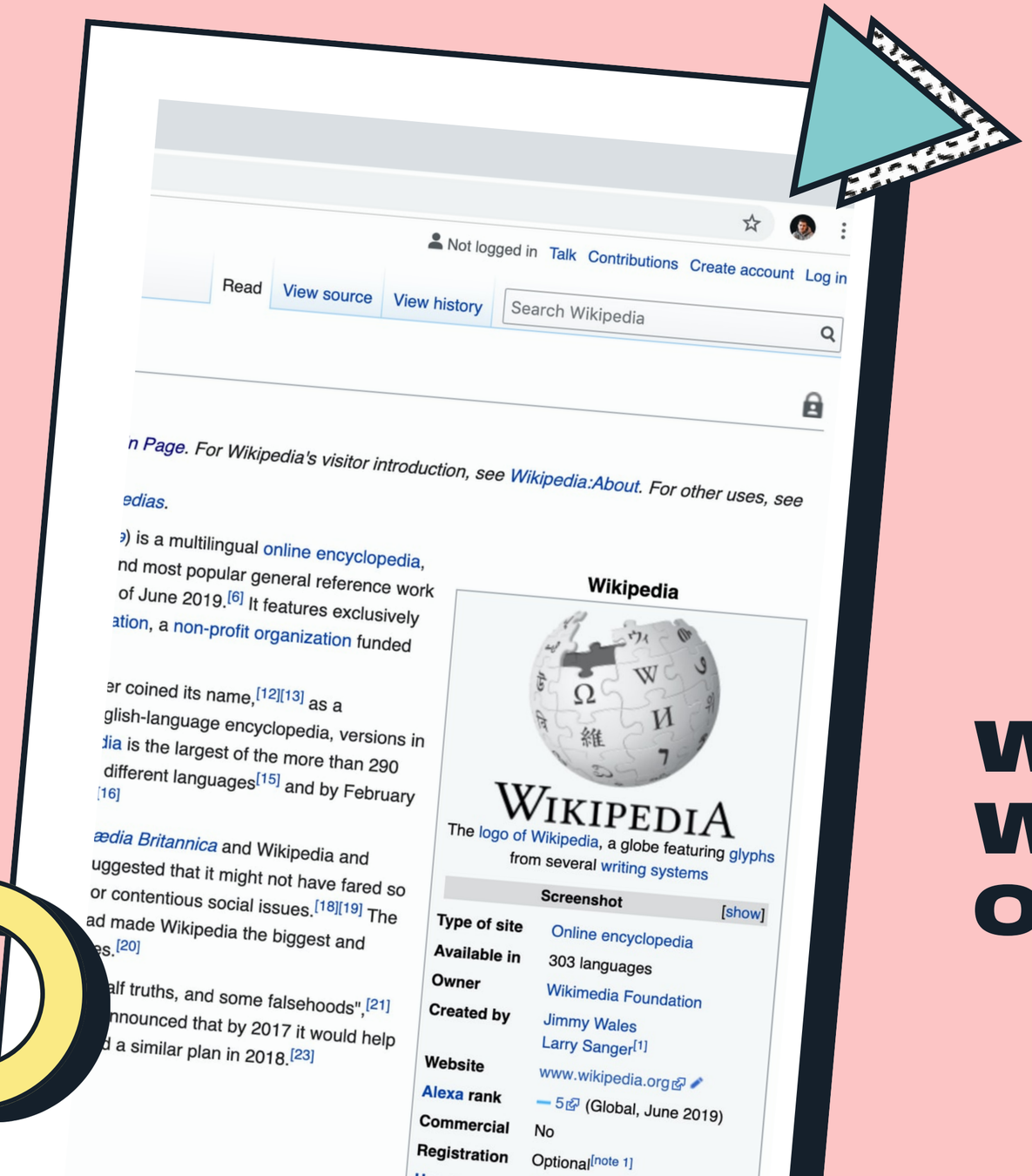


WHY DIDN'T WE CHOOSE ONTOLOGIES?

Pitch

Open Knowledge Graph Creation

in the art domain

Christian Warmuth, Konstantin Dobler, Malte Barth

20.07.21

Knowledge Graphs 2021

Hasso Plattner Institute Potsdam

Pitch

The three steps to knowledge

1

OpenIE

2

Embeddings

3

Clustering

Pitch

1

Open Information Extraction

Pitch

OpenIE

OpenIE extracts triples

- OpenIE6 - using Iterative Grid Labelling and Coordination Analysis
- expects 1 sentence per line - co-reference-resolution beforehand
- unfiltered: ~3 Million triples (with threshold of > 0.8 confidence: ~460.000 triples)
- Bad/Funny Examples:
  - <'It' - 'is' - 'always Picasso'>
  - <'We' - 'learn' - 'more about them and them'>
  - <'One' - 'can skip' - 'truth'>

Pitch

OpenIE

Cleanup afterwards

- Entities smaller than 4 characters
- Relations smaller than 4 characters other than 'was', 'be', 'is'...
- Entity/Relation only numbers/special characters
- Remove pronouns like 'I', 'Them', 'They', 'He' etc.
- ~460.000 triples to 370.000 triples

Pitch

2

EMBEDDINGS

Our Embeddings

- using TransE, ConvE and HolE
- library called Ampligraph
- trained with 100 dimensions
- no pre-trained embeddings as input

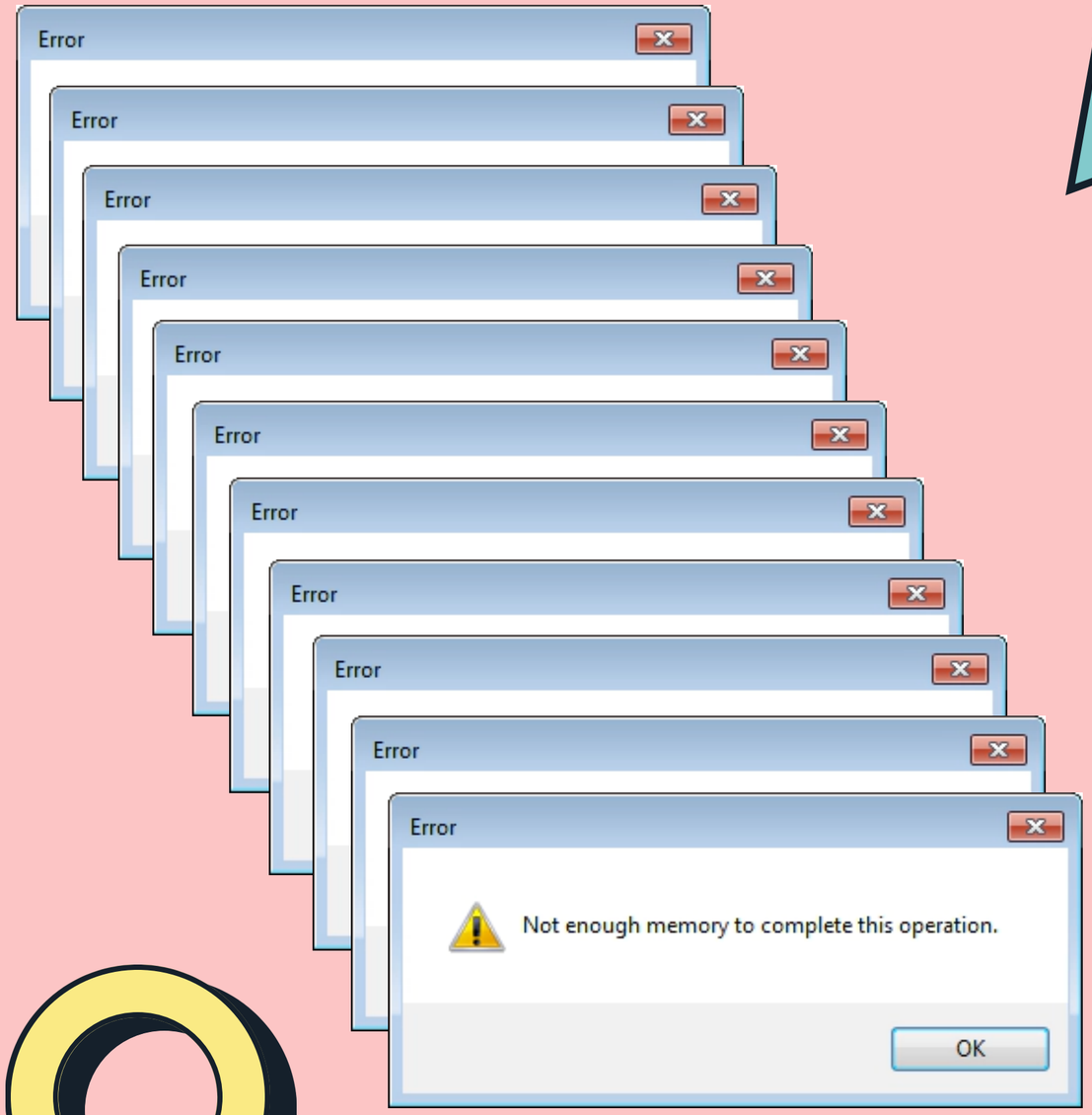
CESI

Our Embeddings

- using TransE, ConvE and HolE
- library called Ampligraph
- trained with 100 dimensions
- no pre-trained embeddings as input

CESI

- using HolE as embedding
- repo from original paper
- trained with 100 dimensions
- using pre-trained GloVE embeddings
- additionally uses some side information



MEMORY ERRORS!!!

Too many dimensions and clustering!

3

Clustering

How to do it at scale?

CESI uses Hierarchical Agglomerative Clustering (HAC)

- HAC is in O(n^2)
- Not feasible for large amount of embeddings (in our case ~500k)
- 'Approximate' variant from CESI paper not (feasibly) usable for custom data

DBSCAN to the rescue!

- Classic 'Density'-based clustering algorithm
- DBSCAN is fast and does not need a predefined number of clusters
- We interpret 'noise' labels as entities / relations without a cluster

Pitch

Canonicalization Results

Input

- ~500k noun phrases
- ~62k relation phrases
- ~370k unique triples

Output

- ~350k noun phrases (difference: ~150k, -30%)
- 13k relation phrases (difference: ~49k, -79%)
- ~345k unique triples (difference: ~25k, -7%)

Pitch

EVALUATION

Pitch

We have no ground truth!

How can we even evaluate our results?

- Create ground truth by manual labelling :(
- Look at the quality of some predefined set of clusters
- Perform a predefined set of queries

Pitch

Evaluation without Ground Truth

Cluster Inspection

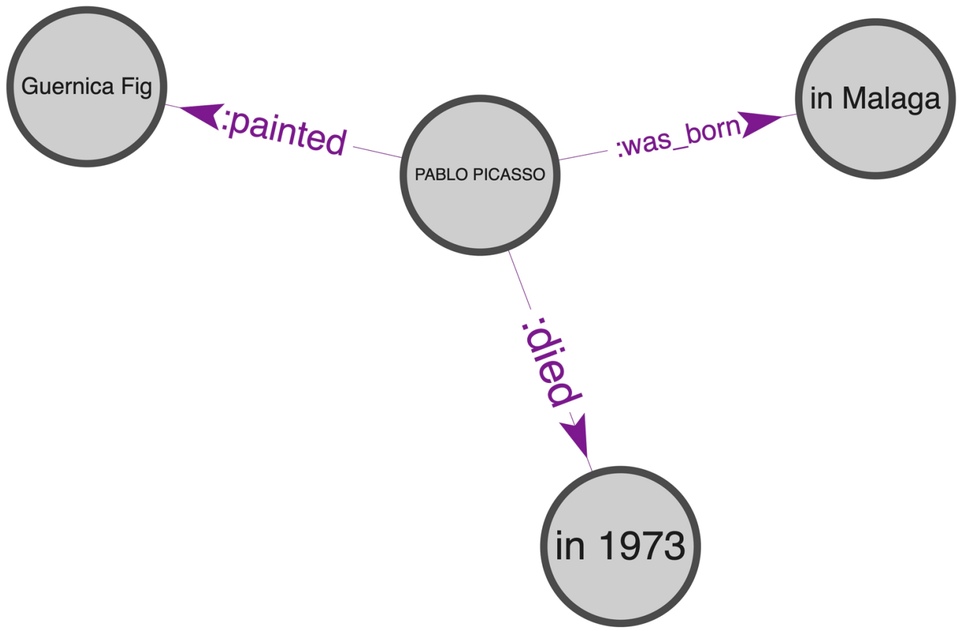
- In general, the cluster quality is just okay

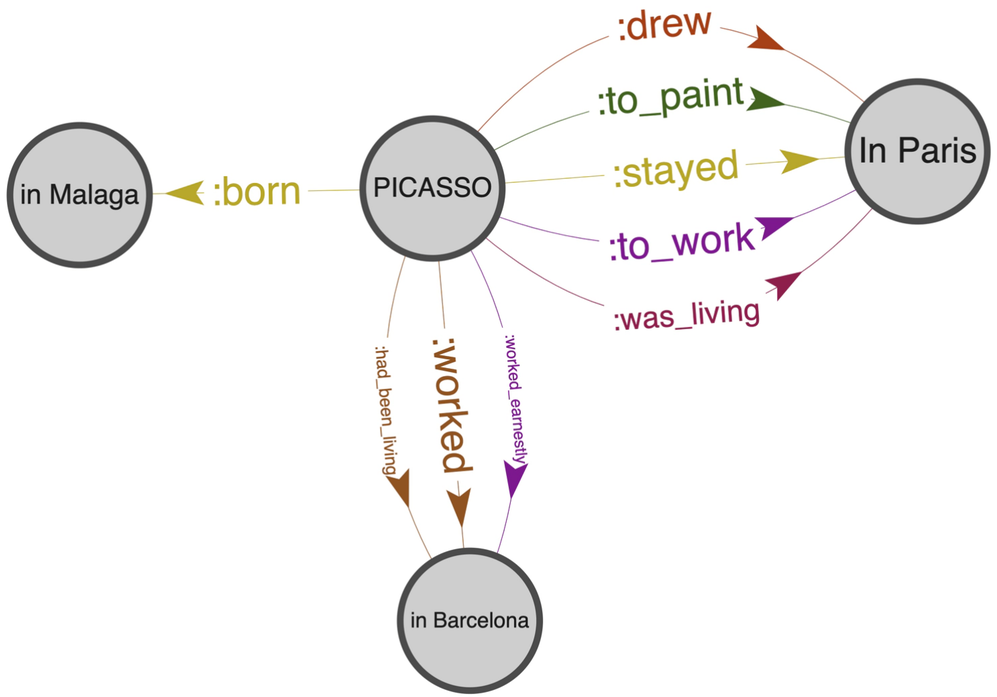
- Some examples for the <Picasso> cluster:
  - <Pablo Picasso>
  - <Most significant artist century>, <Best artist>, ...
  - <upheaval first world war>, <Communist Party>

- CESI seems to be a bit better than the rest

Pitch

LOOKING AT OUR GRAPH





LET'S LOOK AT SOME EXAMPLES

Our example queries

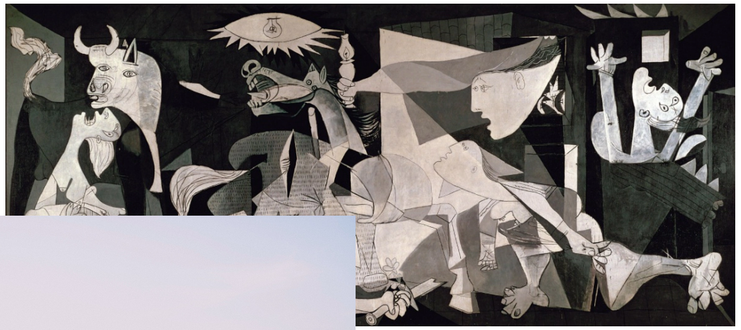
1. Picasso, painted, Guernica

2. Picasso, was_born_in, Malaga

3. Eva Goul, died_of, tuberculosis

4. Olga Picasso, has_son, Paulo Picasso

5. Museu Picasso, in, Barcelona



Pitch

How many matches did we get?

	TransE	ConvE	HoIE	CESI
Query 1	✓	✓	✓	✓
Query 2	✓	✓	✓	✓
Query 3	✗	✗	✓	✓
Query 4	✗	✗	✗	✗
Query 5	✗	✗	✓	✓
Total	**2/5**	**2/5**	**4/5**	**4/5**

Pitch

LEARNINGS

THOROUGH PREPROCESSING IS IMPORTANT

PRE-TRAINED EMBEDDINGS ARE USEFUL

EVALUATION WITHOUT A GROUND TRUTH IS PRETTY HARD

YOU DON'T ALWAYS NEED DOMAIN EXPERTS TO FIND WHAT YOU NEED

Others have seen **what is** and asked **why**. I have seen **what could be** and asked **why not.**

Pablo Picasso