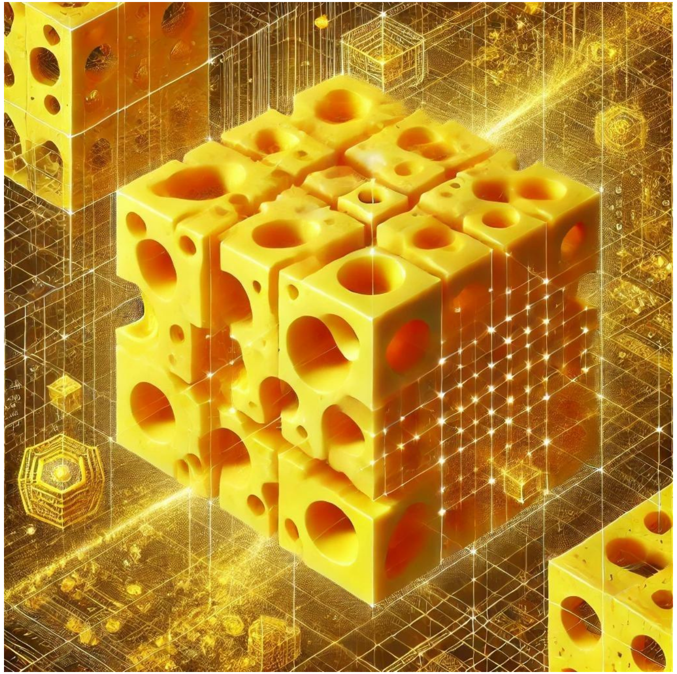
About Search Engines, and how to find relevant answers in High Dimensional Swiss Cheese



What is it about?

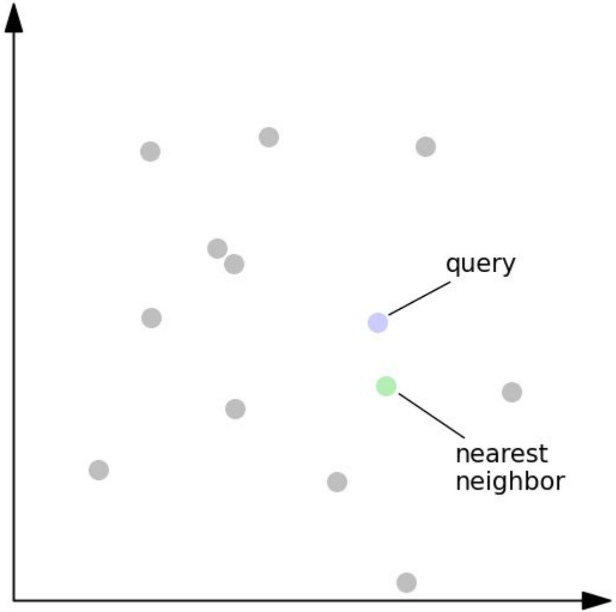
Developing a Relevance Measure for Search Results

- The Relevance Measure shows how relevant the Search Result is with respect to the Question
- I will apply this Relevance Measure to two Search Engines, each fed with a different Book
- We will make n-dimensional Swiss Cheese!

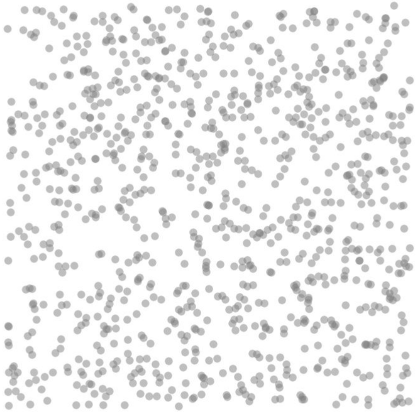
Similarity Search

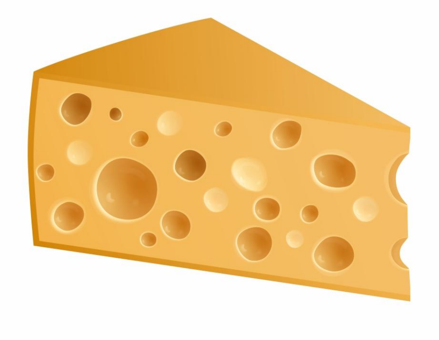
How Search Engines work:

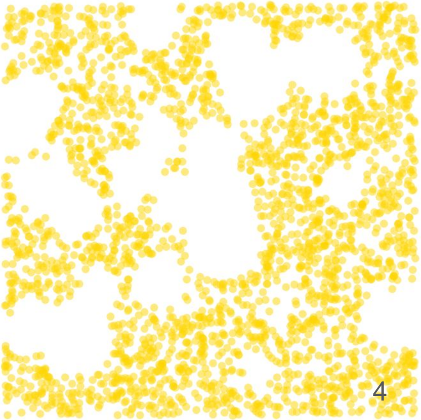
- Information (Corpus, Vector Spaces)
- Pieces of Information (Data Points, Vectors)
- measuring Distances
- Question (Query)
- Nearest Neighbors



What does infinite knowledge look like?
What does any real body of knowledge look like?

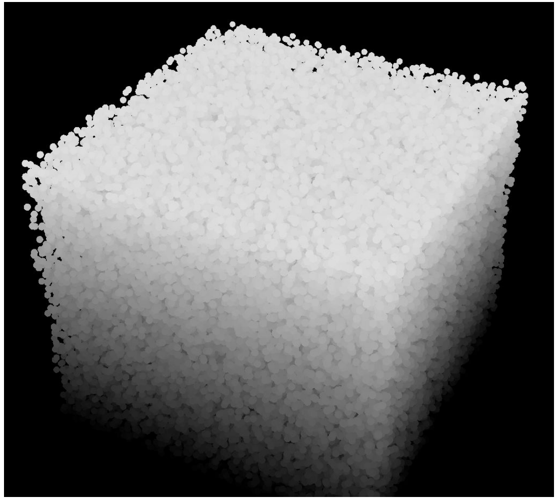


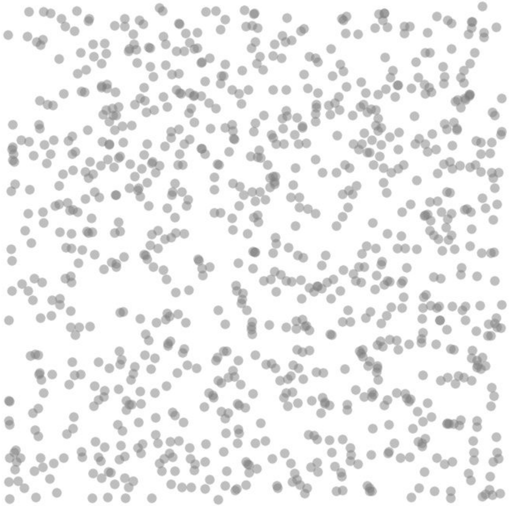


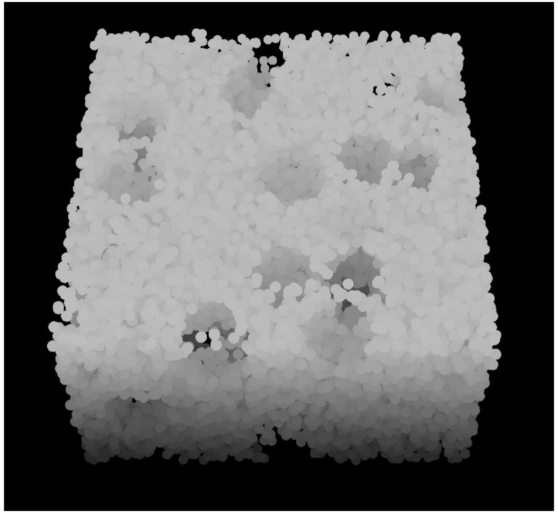


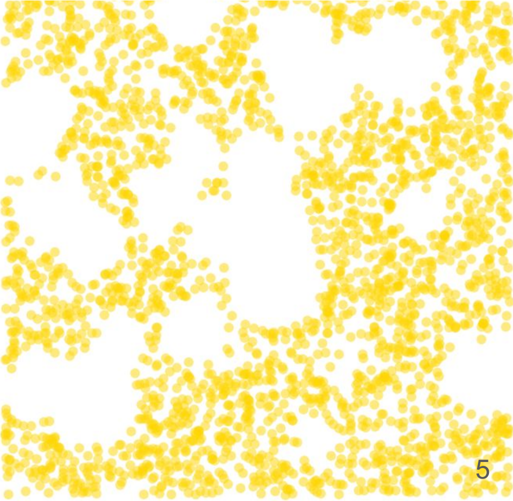
Data - Theoretical

Self generated







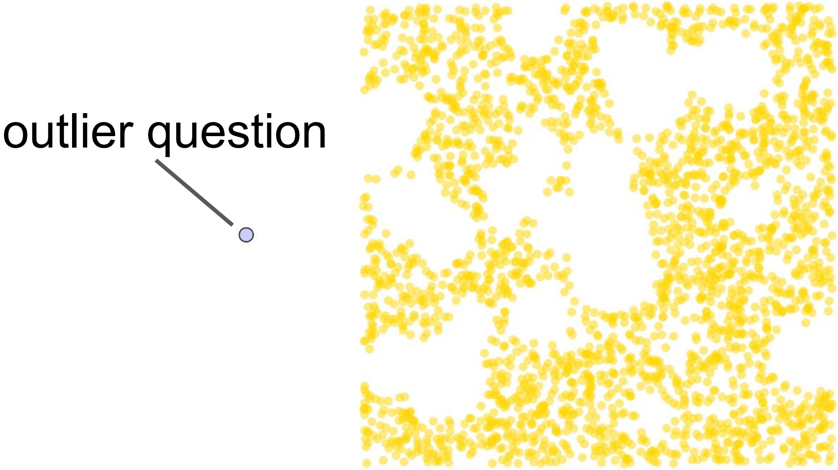


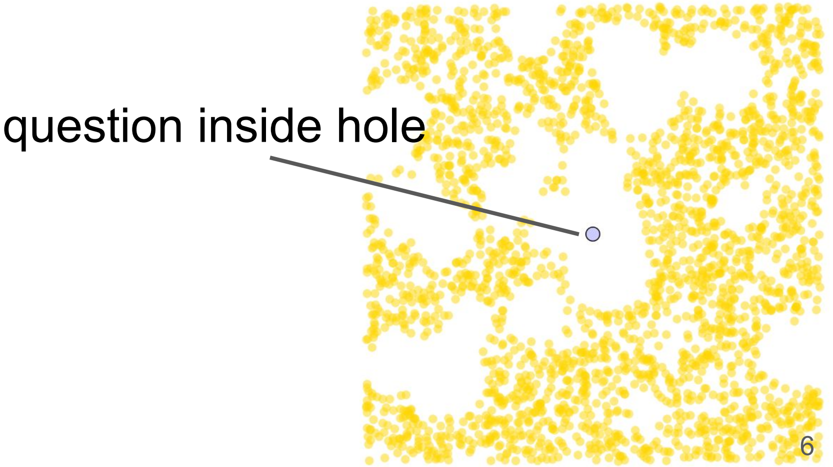
Outliers & Lack of Information

For outlier questions you will still get the best possible answers,

But they are of poor quality because they are not very relevant to the question

We will now develop a way to quantify exactly how relevant the answers are

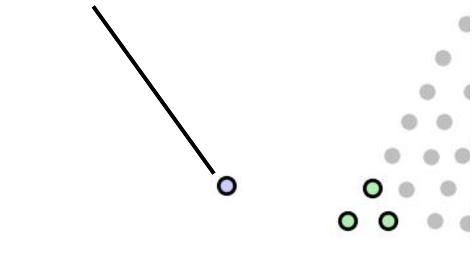




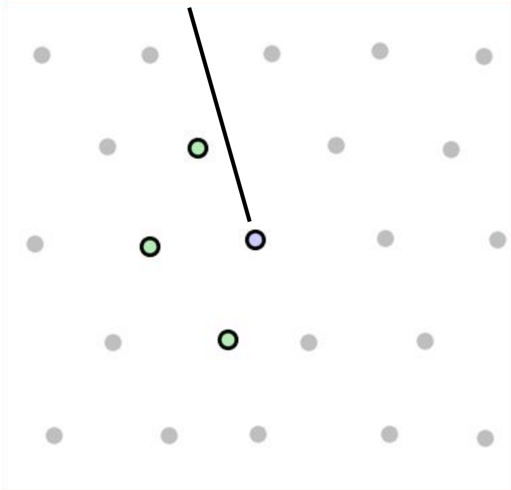
How can we quantify the relevance of Search Results?

- 🟣 question
- 🟢 nearest neighbor (nn)

outlier question

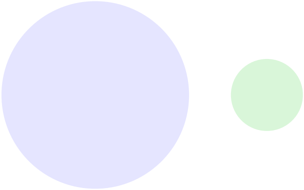


perfectly embedded question

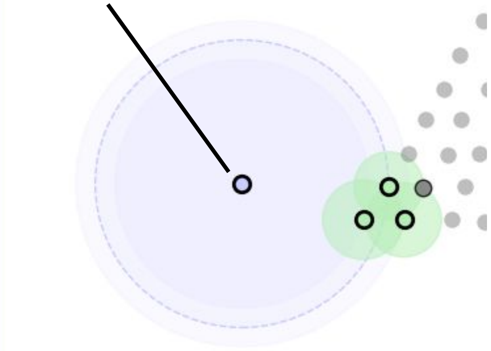


How can we quantify the relevance of Search Results?

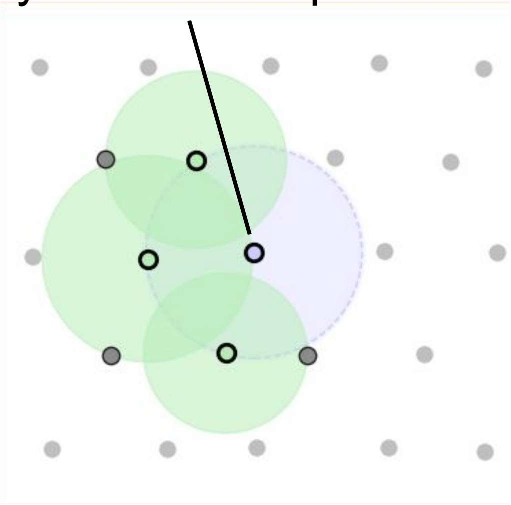
Observe the difference between outliers and embedded questions:

- outlier: 

- embedded:



outlier question



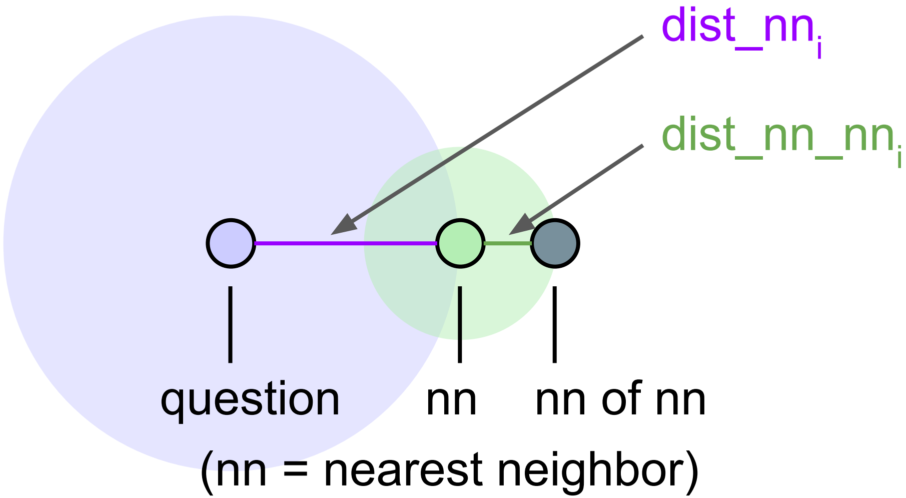
perfectly embedded question

Calculation of Relevance Number

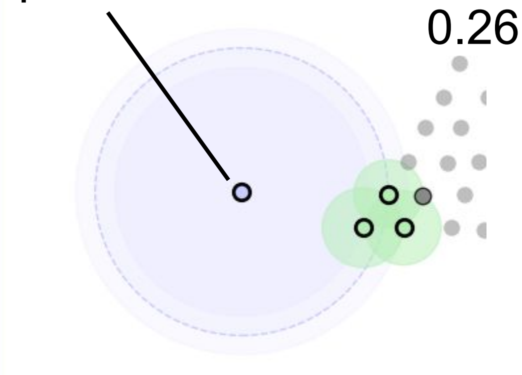
Formula: $\text{relevance} = (1/k) \sum (\text{dist\_nn\_nn}_i / \text{dist\_nn}_i)$

with k = number of search results

Relevance Number is always in interval [0,1]

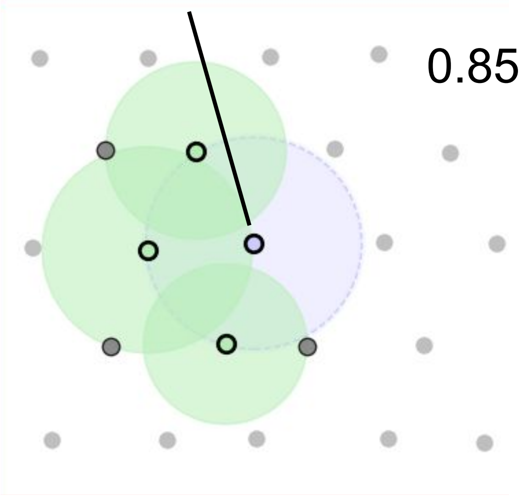


Theoretical Results



outlier question

0.26

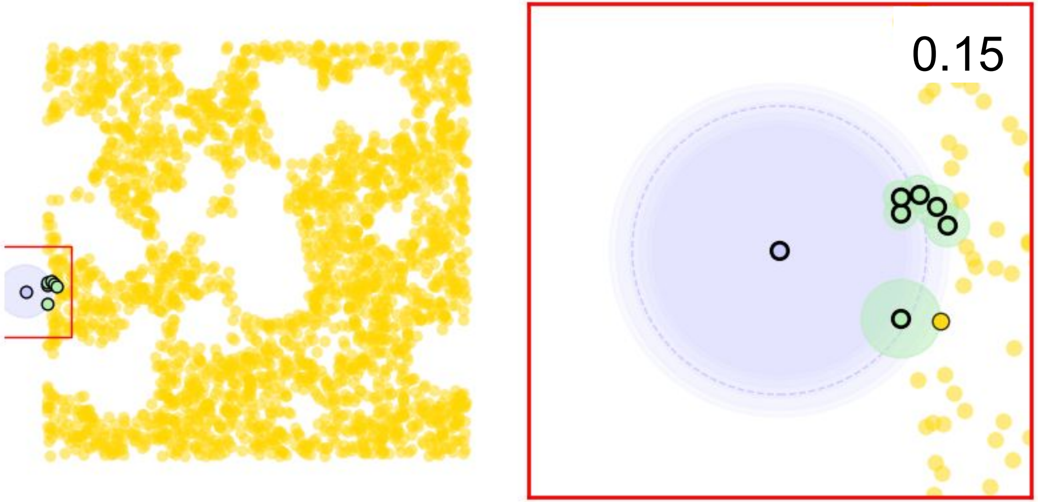


perfectly embedded question

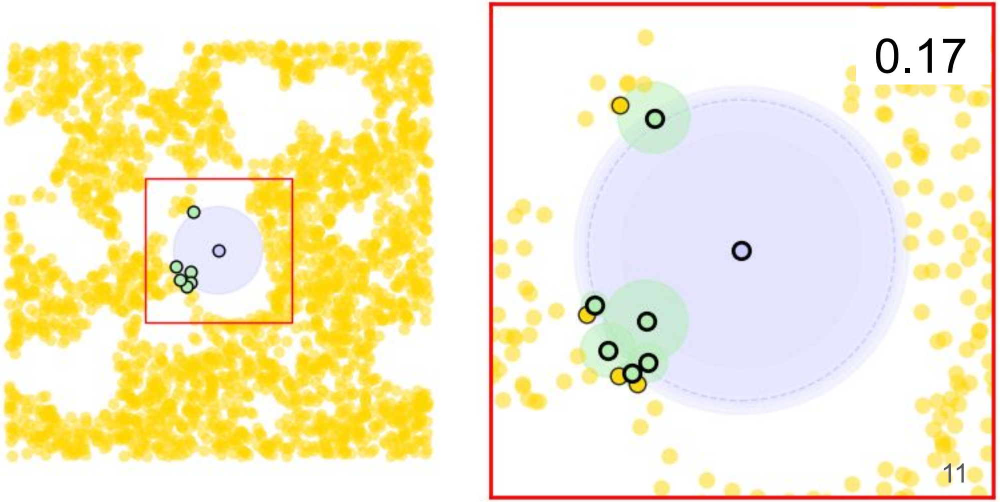
0.85

Theoretical Results

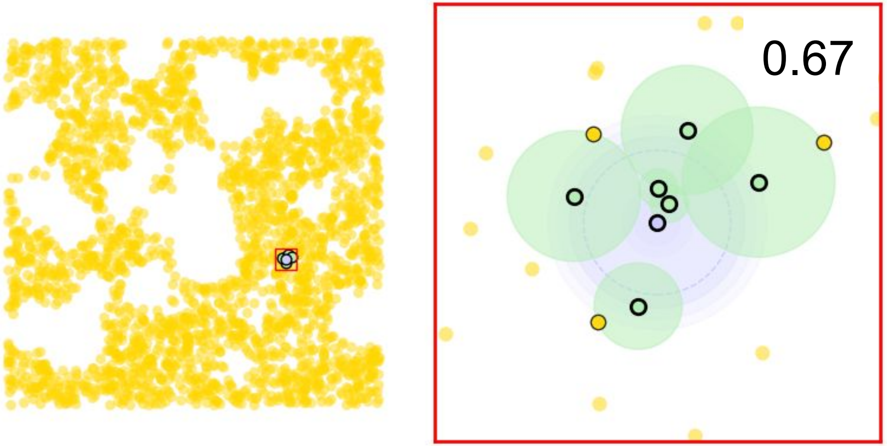
outlier



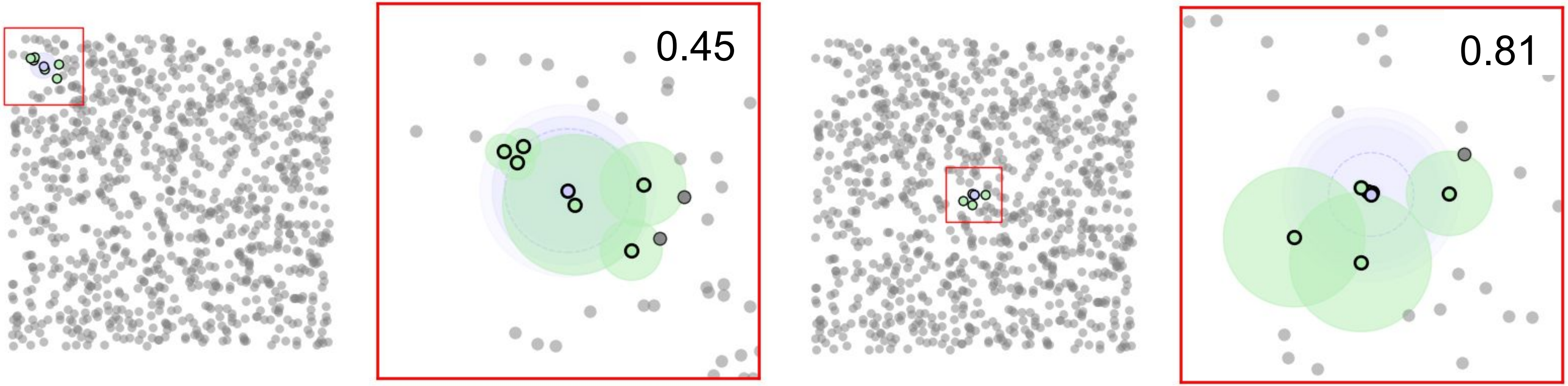
inside hole



Theoretical Results



embedded in the cheese

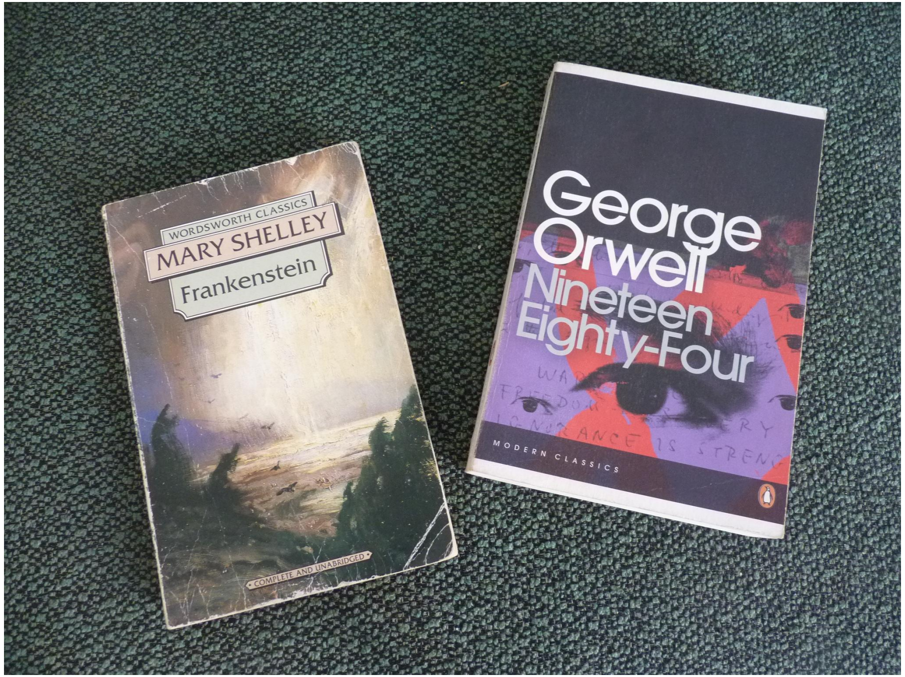


random points

Data - Practical - two Search Engines, two Books

- Search Engine 1: Frankenstein
- Search Engine 2: 1984

source: www.gutenberg.org



Practical Results - two Search Engines, two Books

Questions:

Q1: What did Frankenstein create?

Q2: Who is Henry Clerval?

Q3: What are the three super states?

Q4: What happens in Room 101?

Q5: What courses does WBS Coding School offer?

Practical Results - two Search Engines, two Books

Relevance

Question	Frankenstein	1984
Q1: What did Frankenstein create?	0.73	0.63
Q2: Who is Henry Clerval?	0.77	0.59
Q3: What are the three super states?	0.59	0.77
Q4: What happens in Room 101?	0.54	0.68
Q5: What courses does WBS Coding School offer?	0.47	0.51

Conclusion, Outlook

- We developed a relevance measure for Search Results
- We showed how the relevance measure can be calculated
- We applied the relevance measure to Search Engines with limited domain knowledge

- Having this relevance measure enables us to create trustworthy, honest Search Engines and Chatbots

“Admitting ignorance is wiser than pretending knowledge”

Any Questions?

I'll tell you how relevant they are!