# DIVERSITY AND NOVELTY ON DATA STREAMS

Gökçe Ayduğan                    Melih Baydar

# Content

- Problem Definition
- Motivation / Importance
- Methodology and Expected Result

# Problem Definition

- New Event Detection under a topic

# Content

- Problem Definition
- **Motivation / Importance**
- Methodology and Expected Result

# Redundancy



| ✓ | Name | | Time | Artist |
|---|------|---|------|--------|
| ✓ | Take It Off | | 3:35 | Ke$ha |
| ✓ | We R Who We R | | 3:25 | Ke$ha |

## System Recommends

| Name | ☁ | Time | Artist |
|------|---|------|--------|
| Die Young | | 3:33 | Ke$ha |
| Boots & Boys | | 2:56 | Ke$ha |
| C'mon | | 3:34 | Ke$ha |

# Query Ambiguity

1. genuinely ambiguous queries
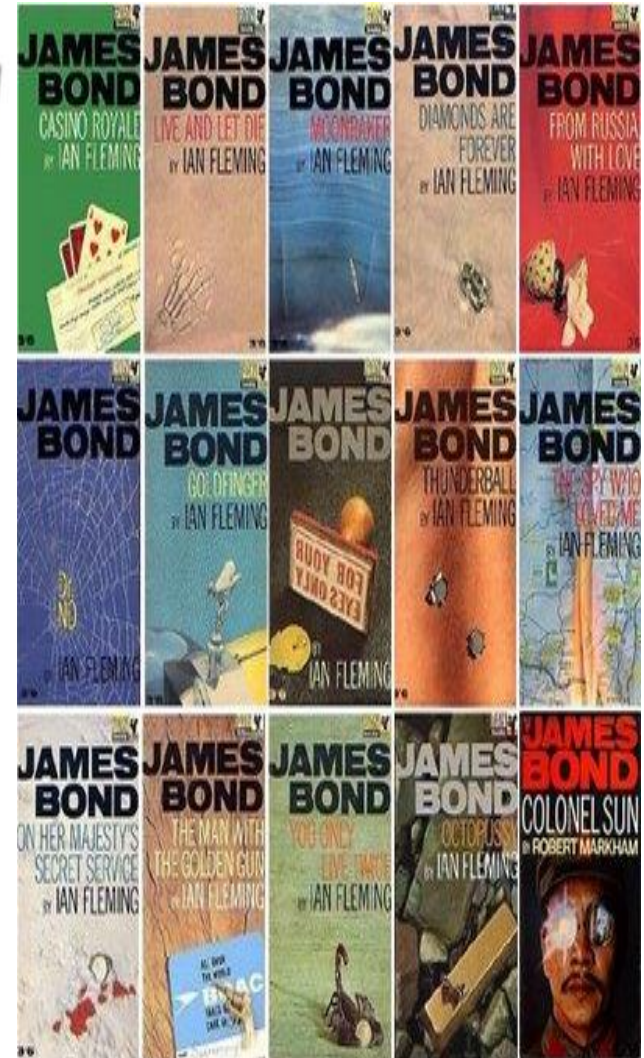2. underspecified queries
3. clear queries

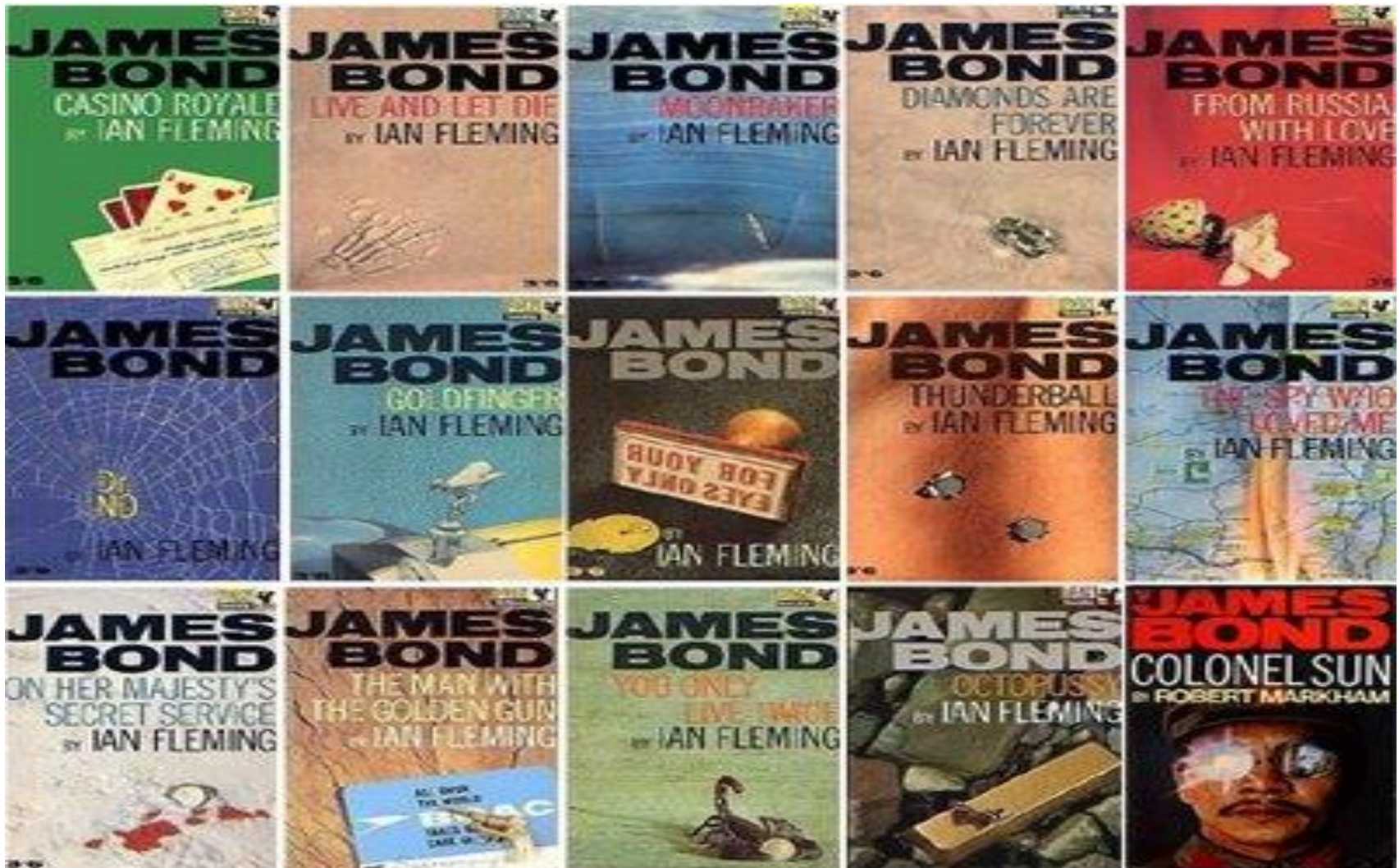# Genuinely ambiguous queries

Query = «bond»
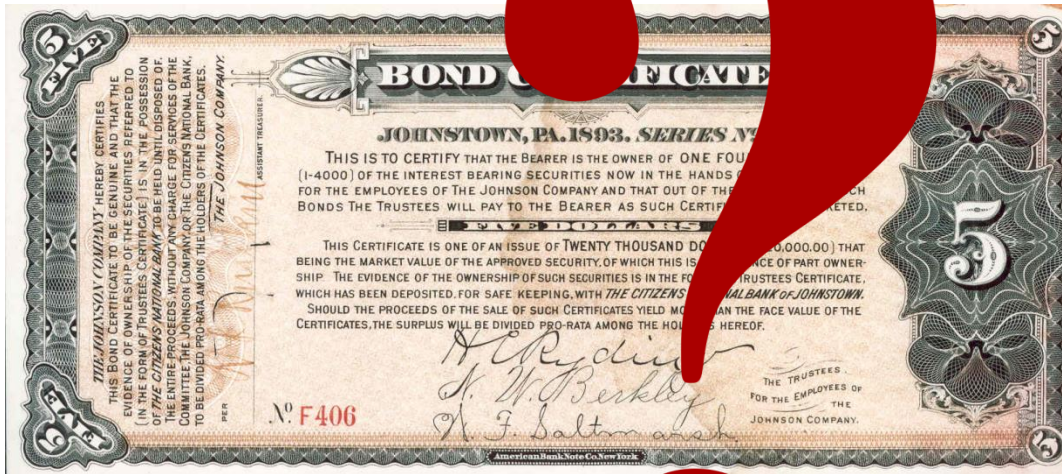
# Underspecified Queries

Query = «James Bond»

# Clear queries

Query = «James Bond books»

# IR Problems

- Redundancy → - Novelty

- Ambiguity → - Diversification

Clarke, Charles LA, et al. "Novelty and diversity in information retrieval evaluation."
Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008.

# Diversification and novelty in Stream Processing

- Analyzing continuous data streams and generating output streams for information consumers

- Aim
  - ✓ information filtering
  - ✓ information aggregation
  - ✓ trend detection
  - ✓ sentence extraction-based summarization

- Data
  - ✓ news articles
  - ✓ intelligence reports
  - ✓ scientific papers
  - ✓ emails
  - ✓ tweets

Santos, Rodrygo Luis Teodoro (2013) Explicit web search result diversification. PhD thesis

# Diversity and Novelty On Data Streams

# Content

- Problem Definition
- Motivation / Importance
- Methodology and Expected Result

# Heuristic Approach

- Preprocess data, calculate probability of each term to exist together with other terms
  - Realize words that are similar or that are relevant.
  - Maybe to enrich tweets
- Create inverted index of tweets
- Get intersection lists for a query
- Cluster result documents
  - Use probabilities in this stage
  - Ideally, clusters contain different topics
- Use LDA(Latent Dirichlet Allocation) to realize cluster representatives
- Mix clusters according to rankings within clusters
- Show results

# References

- Santos, Rodrygo Luis Teodoro (2013) Explicit web search result diversification. PhD thesis

- Ozdemiray, Ahmet Murat, and Ismail Sengor Altingovde. "Explicit search result diversification using score and rank aggregation methods." *Journal of the Association for Information Science and Technology* (2015).

- Rodrygo L. T. Santos,Pablo Castells ,Ismail Sengor Altingovde ,Fazli Can, "Diversity and Novelty on the Web: Search, Recommendation, and Data Streaming Aspects«