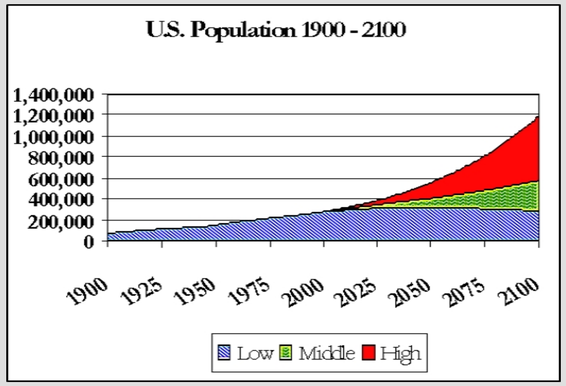
DiagramFlyer: A Search Engine for Data-Driven Diagrams

Shirley Zhe Chen, Michael Cafarella and Eytan Adar
University of Michigan, Ann Arbor

Introduction

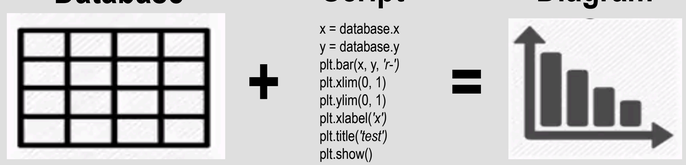
❖**Data-driven diagrams** are an important method for communicating complex information.



❖**Searching for relevant diagrams** would be useful for data-driven professionals.

❖**Existing search systems** can be used to retrieve diagrams.

- ❖Standard **text-based search** engines may be able to retrieve the diagrams' enclosing documents.
- ❖**Image-based search** engines focus on the end product rather than the pipeline that generated the diagram.



DiagramFlyer Demo

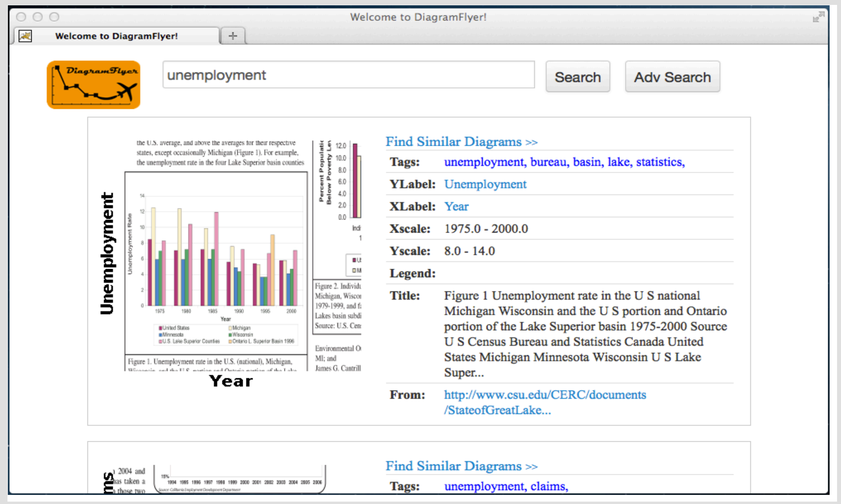
❖DiagramFlyer allows search on data-driven diagrams by **inferring and querying the underlying production pipeline.**

❖DiagramFlyer features search by:

- ❖**Keyword** (traditional keyword queries)
- ❖**Facets** (querying on the pipeline and diagram sub-features)
- ❖**Similarity** (related diagrams)

User Interface and Query Language

❖DiagramFlyer follows the convention of traditional a ten-link search engine result page (SERP) with additional diagram-specific features.

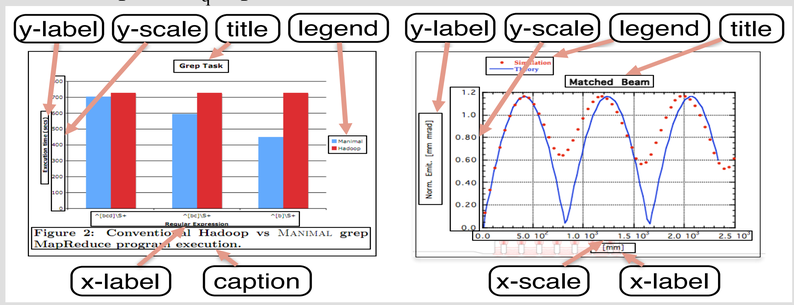


❖**Example 1.** If a user wants to get diagrams about population statistics over the year 1990 to 2014, she can formulate the query as follows:

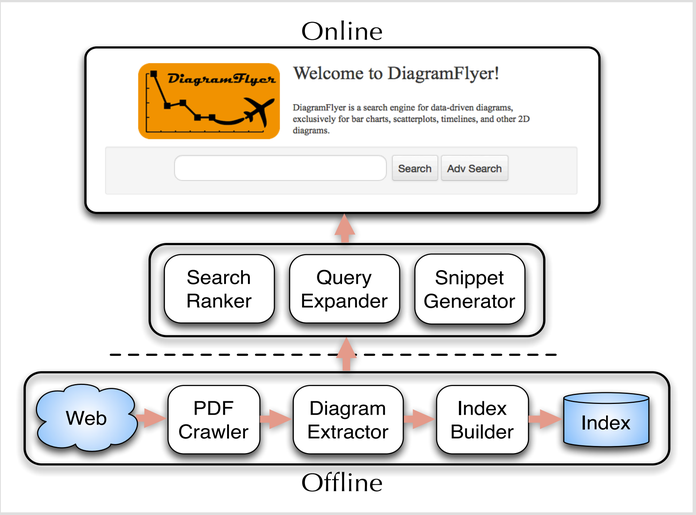
x-label: year AND
x-scale: from: 1990 to: 2014 AND
y-label: population

❖**Example 2**. After finding a relevant diagram d_q, a user may want to retrieve all the relevant diagrams that could potentially be generated from the same underlying dataset. The query can be formulated as:

x-label:expand(d_q.x-label) AND
y-label:expand(d_q.y-label) AND
title: d_q.title AND
caption: d_q.caption



DiagramFlyer Framework



❖The **offline** stage crawls and indexes diagrams by:

- ❖**PDF Crawler** (finds 153k PDFs on .edu websites).
- ❖**Diagram Extractor** (extracts 319k diagrams with images and metadata).
- ❖**Index Builder** (indexes the text and metadata).

❖The **online** system supports dynamic access and search through:

- ❖**Search Ranker** (computes relevance, ranking from keywords).
- ❖**Query Expander** (expands keywords by using related terms found from mining 14 million HTML lists).
- ❖**Snippet Generator** (provides visual summary of results in an easy-to-read manner).

References

[1] Zhe Chen, Michael Cafarella, Eytan Adar, DiagramFlyer: A Search Engine for Data-Driven Diagrams, WWW Demo, Florence, Italy, 2015

[2] Shirley Zhe Chen, Michael J. Cafarella, and Eytan Adar, Searching for Statistical Diagrams, In Frontiers of Engineering, National Academy of Engineering, Washington DC, 2012.

[3] YouTube Demo Video: http://youtu.be/B7I1_o23N38