Seeking Order in Disorder: Towards Accurate and Efficient Document Analytics

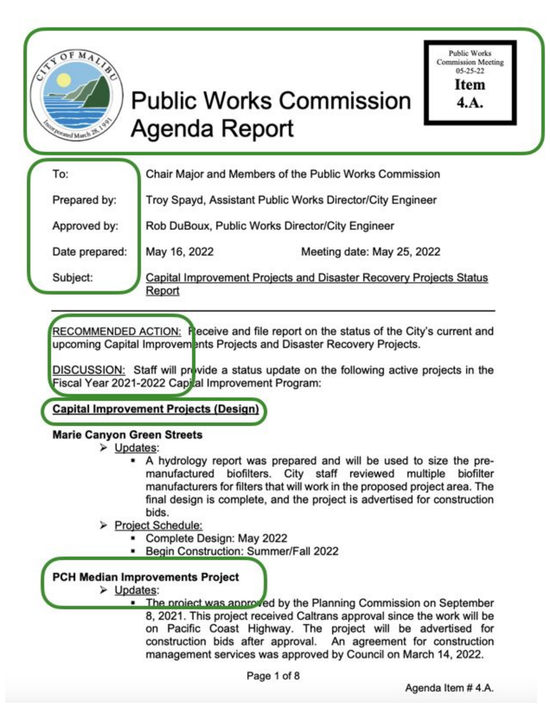
Yiming Lin[1], Madelon Hulsebos[1], Ruiying Ma[2], Shreya Shankar[1],
Sepanta Zeighami[1], Eugene Wu[3], Aditya Parameswaran[1]

[1]*UC, Berkeley,* [2]*Tsinghua University,* [3]*Columbia University*

Background & Question

- Over 80% of data exists in unstructured formats, and extracting values from unstructured documents remains a considerable challenge.
- LLMs provide us the ability to extract semantic objects, labels, and relationships from text much easier than before
- Q: *can we build a data management system for text data and expose a SQL-like query language for perform advanced analytics beyond simple retrieval?*

Templatized Documents



Public Works Commission Agenda Report

Public Works Commission Meeting 05-25-22 Item 4.A.

To: Chair Major and Members of the Public Works Commission
Prepared by: Troy Spayd, Assistant Public Works Director/City Engineer
Approved by: Rob DuBoux, Public Works Director/City Engineer
Date prepared: May 16, 2022 Meeting date: May 25, 2022
Subject: Capital Improvement Projects and Disaster Recovery Projects Status Report

RECOMMENDED ACTION: Receive and file report on the status of the City's current and upcoming Capital Improvements Projects and Disaster Recovery Projects.

DISCUSSION: Staff will provide a status update on the following active projects in the Fiscal Year 2021-2022 Capital Improvement Program:

Capital Improvement Projects (Design)

Marie Canyon Green Streets
- Updates:
  - A hydrology report was prepared and will be used to size the pre-manufactured biofilters. City staff reviewed multiple biofilter manufacturers for filters that will work in the proposed project area. The final design is complete, and the project is advertised for construction bids.
- Project Schedule:
  - Complete Design: May 2022
  - Begin Construction: Summer/Fall 2022

PCH Median Improvements Project
- Updates:
  - The project was approved by the Planning Commission on September 8, 2021. This project received Caltrans approval since the work will be on Pacific Coast Highway. The project will be advertised for construction bids after approval. An agreement for construction management services was approved by Council on March 14, 2022.

Page 1 of 8 Agenda Item # 4.A.

- Project Schedule:
  - Complete Design: March 2022
  - Advertise: Spring/Summer 2022
  - Begin Construction: Summer/Fall 2022

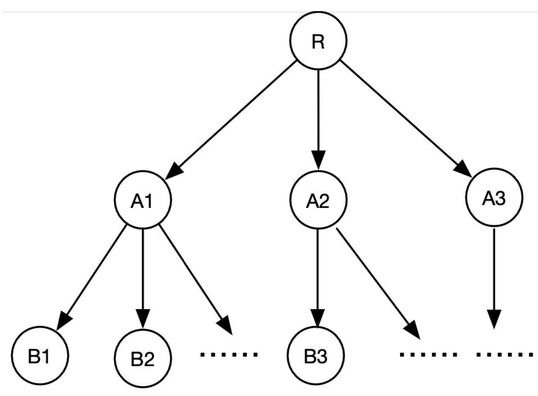
PCH Signal Synchronization System Improvements Project
- Updates:
  - This project will be presented to the Planning Commission in May 2022. This project requires Caltrans approval since the work will be on PCH. The project reports and plans are being routed through Caltrans for final approval. It is anticipated that the project will have final approval by June 2022. The project will be advertised for construction bids shortly after final approval. If possible, the construction of this project will begin in conjunction with the PCH Median Improvement
- Project Schedule:
  - Complete Final Design: Spring 2022
  - Advertise: Summer 2022
  - Award Contract and Begin Construction: Summer/Fall 2022

Westward Beach Road Repair Project
- Updates:
  - On March 28, 2022, The City Council modified the scope of work for this project to include the repair of the existing road as necessary. The proposed revision makes the Measure M funding not eligible for this project. The revised project budget will be included in the Fiscal Year 2022-2023 Budget. Upon the adoption of the budget, the revised project will be redesigned and constructed.
- Project Schedule:
  - Complete Design: Winter 2022
  - Begin Construction: Winter 2022/Spring 2023

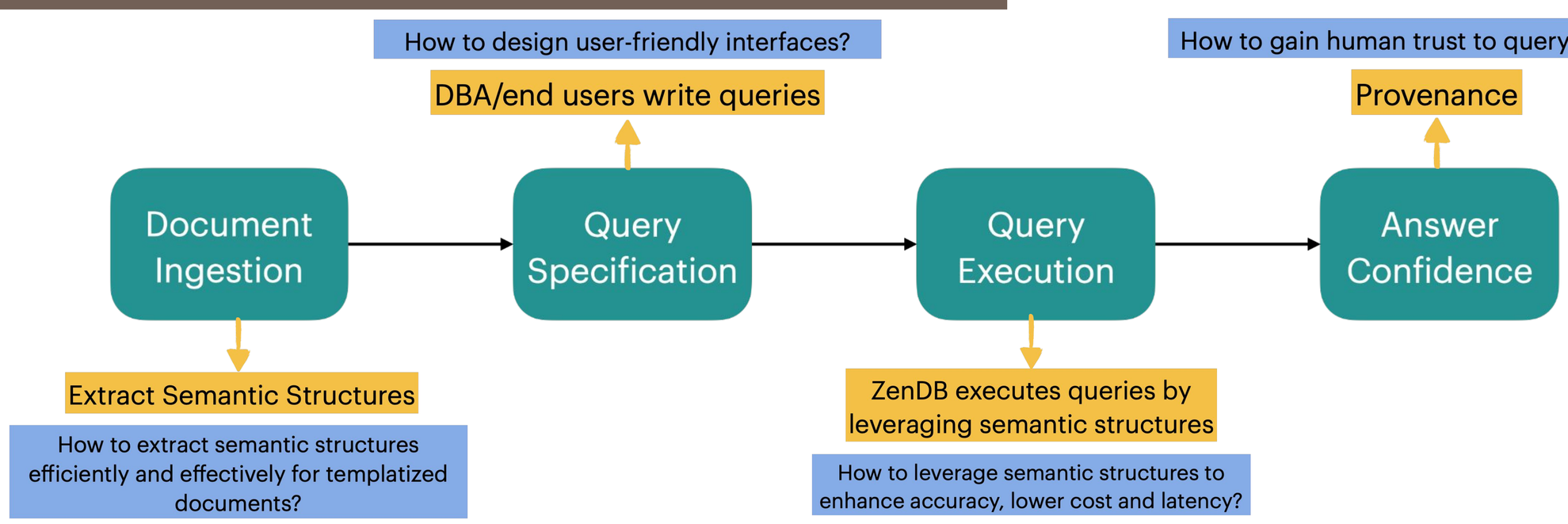
Civic Center Water Treatment Facility Phase 2
- Updates:
  - Individual letters were mailed to all properties within Phase 2 with their preliminary estimated assessments in July 2021. Staff has been communicating with the property owners regarding their proposed assessments.
  - The MOU has been amended modifying the deadline to the formation of the assessment district to June 30, 2022. A new request for further modification of the schedule has been requested.
  - Staff mailed easement documents to property owners for review and execution in July and has followed up with an additional letter to those property owners.
- Project Schedule:
  - Complete Design: December 2021

Page 2 of 8 Agenda Item # 4.A.

R: Public Works Commission Agenda Report
- A1: Capital Improvement and Disaster Recovery Projects (Design)
  - B1: 2022 Morning View Resurfacing & Storm Drain Improvements
  - B2: PCH Median Improvements Project
  - .....
- A2: Street Construction Projects (Construction)
  - B3: Malibu Canyon Road Traffic Study
  - .....
- A3: Disaster Projects (Completed)
- .....



ZenDB's Workflow



Query Interfaces

- Require minimal domain knowledge:
  - table name and description
  - Attribute name/type/description

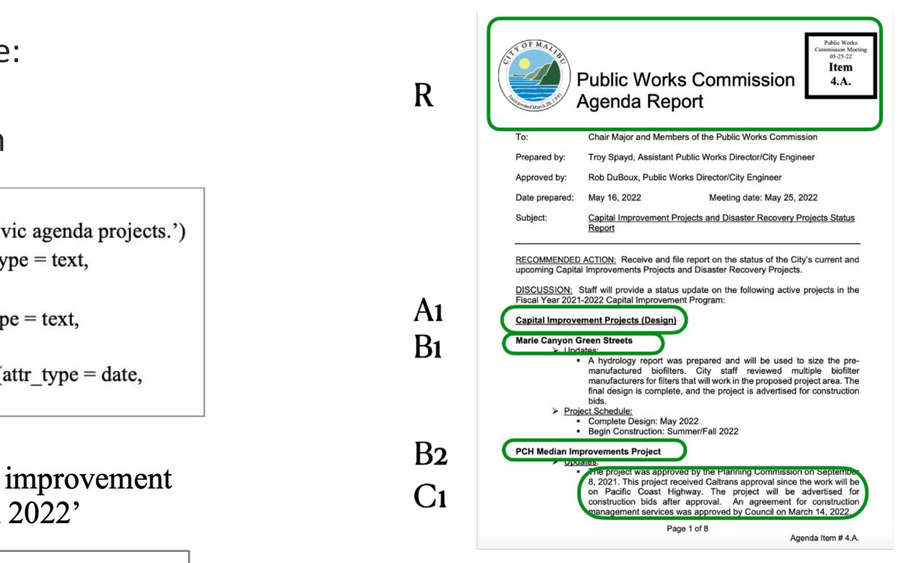
CREATE TABLE Projects **AS** (table_desp =
'The projects table contains the desciption of a set of civic agenda projects.')
CREATE ATTRIBUTE name **ON** Projects **AS** (attr_type = text,
attr_desp = 'name of project')
CREATE ATTRIBUTE type **ON** Projects **AS** (attr_type = text,
attr_desp = 'type of project')
CREATE ATTRIBUTE begin_time **ON** Projects **AS** (attr_type = date,
attr_desp = 'begin time of project')

Q1: 'What is the number of capital improvement projects that start later than 2022'

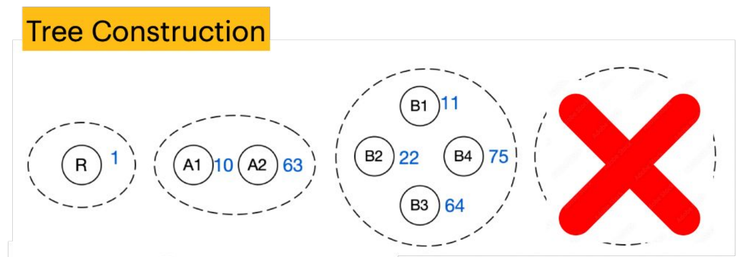
Q2:
```
SELECT COUNT(Projects.name)
FROM Projects,
WHERE Projects.type = 'Capital Improvement'
AND Projects.begin_time > '2022'
```

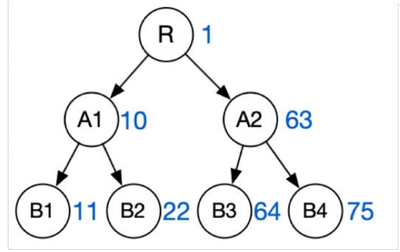
Sketch for each node:
- Name of current node and ancestors
- Summary
- Top-1 sentence similar to query

Semantic Structure



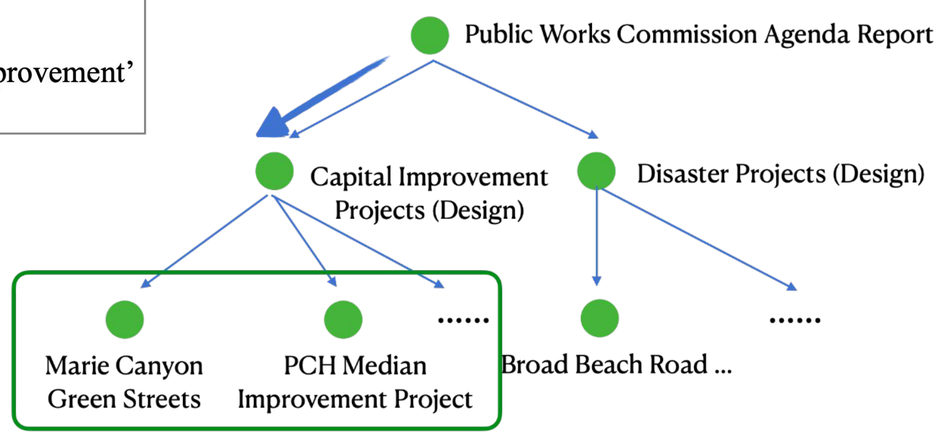
Tree Construction





- Pre-order = reading order
- Nodes in same cluster = same level

Documents sharing templates - by **no-LLM** visual patterns matching



Experiment Takeaways

- **ZenDB VS LLM** (all GPT-4)
  - Up to **+29%** precision, **+31%** recall, **30x** saving of costs, **4x** latency saving
- **ZenDB VS RAG** (all GPT-4)
  - Up to **+61%** precision, **+80%** recall, **1.7x** higher cost and **1.3x** higher latency
- **ZenDB** + **GPT-3.5-Turbo (100x cheaper)**
  - Paying **one dollar**, you can run **5.5k SQLs** on average on single document, with usable quality (**-7%** precision and **-5%** recall VS ZenDB + GPT-4)