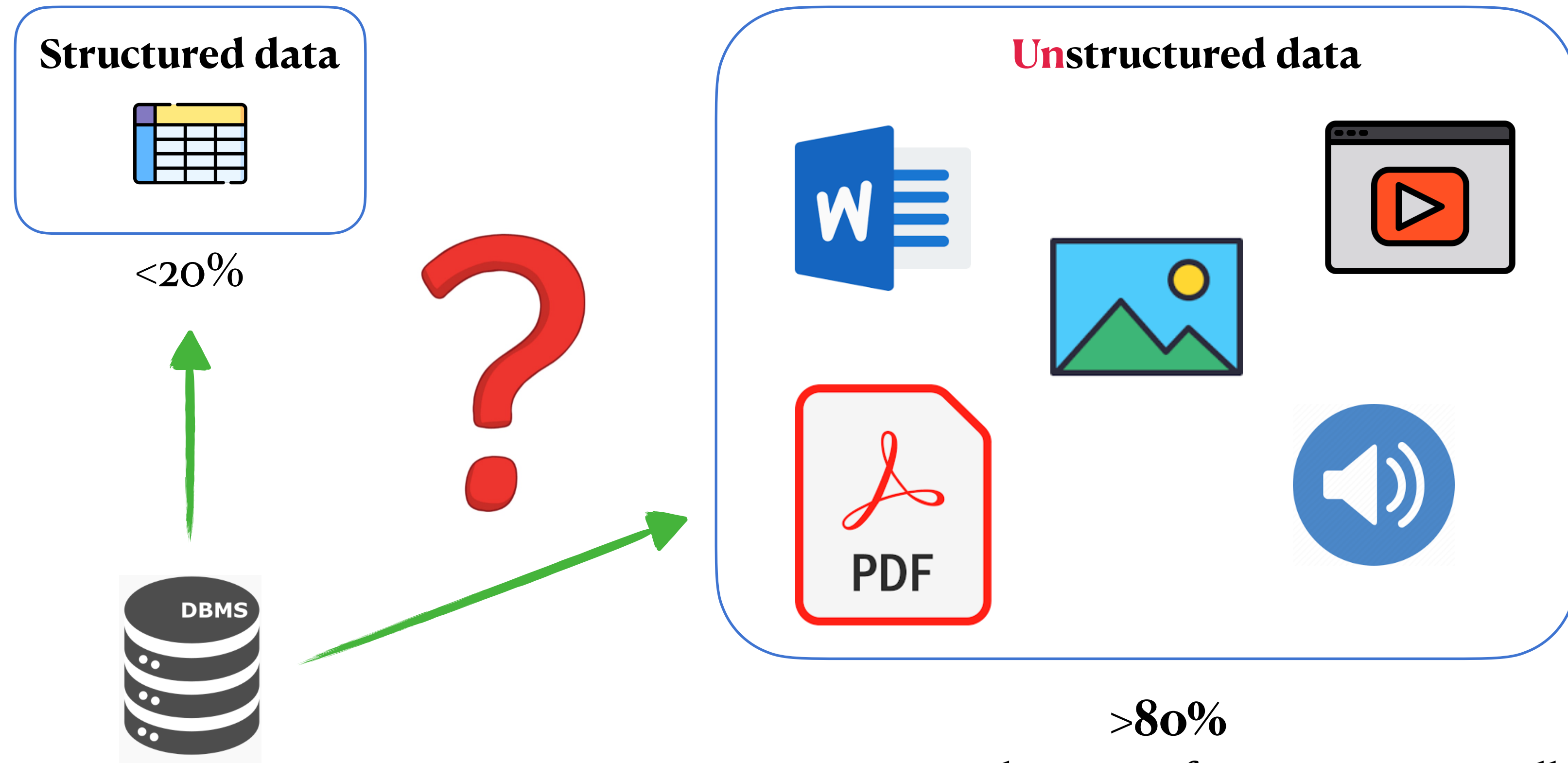


Seeking Order in Disorder: Towards Accurate and Efficient Document Analytics

Yiming Lin, Madelon Hulsebos, Ruiying Ma, Shreya Shankar,
Sepanta Zeighami, Aditya Parameswaran, Eugene Wu

Unstructured Data Management



Growing at the rate of over 50% annually


Can we successfully query or extract value from **unstructured documents**?

Can we build a database system for unstructured documents?

Analyzing Unstructured Documents is Hard

They are free form, have heterogeneous visual formats and layouts, potentially long and complex...

To:
Prepared by:
Approved by:
Date prepared:
Subject:



**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:
upcoming Capital Improver

DISCUSSION: Staff will p
Fiscal Year 2021-2022 Ca

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.

- > **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

- > **Project Schedule:**
 - Complete Final Design: Spring 2022
 - Advertise: Summer 2022

Civic Center Water Treatment Facility Phase 2

- > **Updates:**

Civic Agenda Report from Malibu City
collected by collaborators
from Big Local News at Stanford

Analyzing Unstructured Documents is Hard

They are free form, have heterogeneous visual formats and layouts, potentially long and complex...

**NOTICE OF PROBABLE VIOLATION
PROPOSED CIVIL PENALTY
and
PROPOSED COMPLIANCE ORDER**

VIA ELECTRONIC MAIL TO: dllamp@cvrenergy.com; kakuehn@cvrenergy.com;
brecord@cvrenergy.com; rfmcgill@cvrenergy.com

June 2, 2023

David Lamp
President-Crude Transportation
Coffeyville Resources Crude Transportation, LLC
P.O. Box 3516
411 N.E. Washington Boulevard
Bartlesville, Oklahoma 74006

CPF 3-2023-008-NOPV

Dear Mr. Lamp:

From August 23, 2021 to October 4, 2021, a representative of the Pipeline and Hazardous Materials Safety Administration (PHMSA), Office of Pipeline Safety (OPS), pursuant to Chapter 601 of 49 United States Code (U.S.C.), inspected Coffeyville Resources Crude Transportation, LLC's (CRCT) Hazardous Liquids Crude Pipeline System in Oklahoma and Kansas.

As a result of the inspection, it is alleged that CRCT has committed probable violations of the Pipeline Safety Regulations, Title 49, Code of Federal Regulations (CFR). The items inspected and the probable violations are:

- § 195.412 Inspection of rights-of-way and crossings under navigable waters.**
(a) Each Operator shall, at intervals not exceeding 3 weeks, but at least 26 times each calendar year, inspect the surface conditions on or adjacent to each pipeline

right-of-way. Methods of inspection include walking, driving, flying or other appropriate means of traversing the right-of-way.

CRCT failed to satisfy the requirements of § 195.412(a) by not using an appropriate method for inspection of pipeline right-of-way. During the field inspection at the Hooser-Broome 8" pipeline segment west of Bee Creek Valve (Lat. 37.053019, Long. -95.966233), PHMSA observed that CRCT failed to adequately clear the right-of-way of tree cover, thereby preventing effective aerial patrolling. The right-of-way was covered by a dense tree canopy extending approximately 1,200 feet, which prevented details of the surface conditions on and adjacent to each pipeline right-of-way from being observed. An appropriate means of inspecting the right-of-way in light of these obstructions, such as walking or driving along the right-of-way, was not performed. Therefore, CRCT is in violation of § 195.412(a).

- § 195.573 What must I do to monitor external corrosion?**
(a) *Protected pipelines.* You must do the following to determine whether cathodic protection required by this subpart complies with § 195.571:
(1) Conduct tests on the protected pipeline at least once each calendar year, but with intervals not exceeding 15 months. However, if tests at those intervals are impractical for separately protected short sections of bare or ineffectively coated pipelines, testing may be done at least once every 3 calendar years, but with intervals not exceeding 39 months.

CRCT failed to satisfy the requirements of § 195.573(a)(1) by not performing cathodic protection testing on the protected Shidler pipeline segment within the required interval of at least once each calendar year, but not exceeding 15 months.

From a review of CRCT's records, PHMSA found that the Shidler cathodic protection test surveys were conducted on 2/28/18 and 6/17/19, which exceeded the allowable 15-month interval by 20 days. Therefore, CRCT is in violation of § 195.573(a)(1). This violation was issued as a Warning in CPF #320195021, Item #2.

CRCT did not demonstrate that tests at least once each calendar year, but with intervals not exceeding 15 months, were impractical for the affected pipeline segment.

- § 195.573 What must I do to monitor external corrosion?**
(a)
(e) *Corrective action.* You must correct any identified deficiency in corrosion control as required by § 195.401(b). However, if the deficiency involves a pipeline in an integrity management program under § 195.452, you must correct the deficiency as required by § 195.452(h).

CRCT failed to correct identified deficiencies in its corrosion control. From the inspection of corrosion control records, PHMSA found that CRCT's inspections in calendar years 2020 and 2021 for four of its steel breakout tanks showed that the minimum protection criteria of NACE SP 0169 was not met as required by § 195.571.

If all we have are only these raw documents, What can we do?

Notice of Violation about
Hazardous Materials Safety Administration
from US Department of Transportation

What Types of Analyses are Feasible?

Journalist: give me the number of projects related to **Capital Improvement starting after 2021.**



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.

- > 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

Capital Improvement
Projects (Design)

Begin Construction:
Summer/Fall 2022

Marie Canyon Green Streets

How Can We Answer the Query? - Attempt 1: LLM

Journalist: give me the number of projects related to **Capital Improvement** **starting after 2021**.

- Prompt: Question + Complete Document
- How does it work?
 - LLM: GPT-4-32k
 - Documents: 40 Civic Agenda Reports

LLM gives correct answers on **41%** of documents,
taking **12.2\$ for one question**, spending **3 min...**


How Can We Answer the Query? - Attempt 2: RAG

Journalist: give me the number of projects related to **Capital Improvement starting after 2021.**

To reduce the cost...

- Chunk the document
- Return Top-k chunks that are most similar to query
- Prompt: query + top-k chunks
- How does it work?
- Correct on **3%** documents...
- Why?

Capital Improvement Projects (Design)



**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.

Westward Beach Road Repair project

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

Westward Beach Road Repair Project

Project Schedule:
Complete Final Design: Spring 2022
Advertise: Summer 2022
Award Contract and Begin Construction: Summer/Fall 2022

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.

Takeaways from LLM and RAG Attempts

LLM:

- **High cost** on large document collections
- **Undesirable accuracy** for long context - “Lost in the middle” [1,2]

• RAG:

- **Cheaper** but **inaccurate** - miss the right text portions due to reliance on physical chunking


Any other hope to do it better?

[1] Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *Transactions of the Association for Computational Linguistics* 12 (2024): 157-173.

[2] Bai, Yushi, et al. "Longbench: A bilingual, multitask benchmark for long context understanding." *arXiv preprint arXiv:2308.14508* (2023).

Structure In an UnStructured World?

Unstructured Documents are often **semantically structured!**



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

Public Works Commission Agenda Report

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

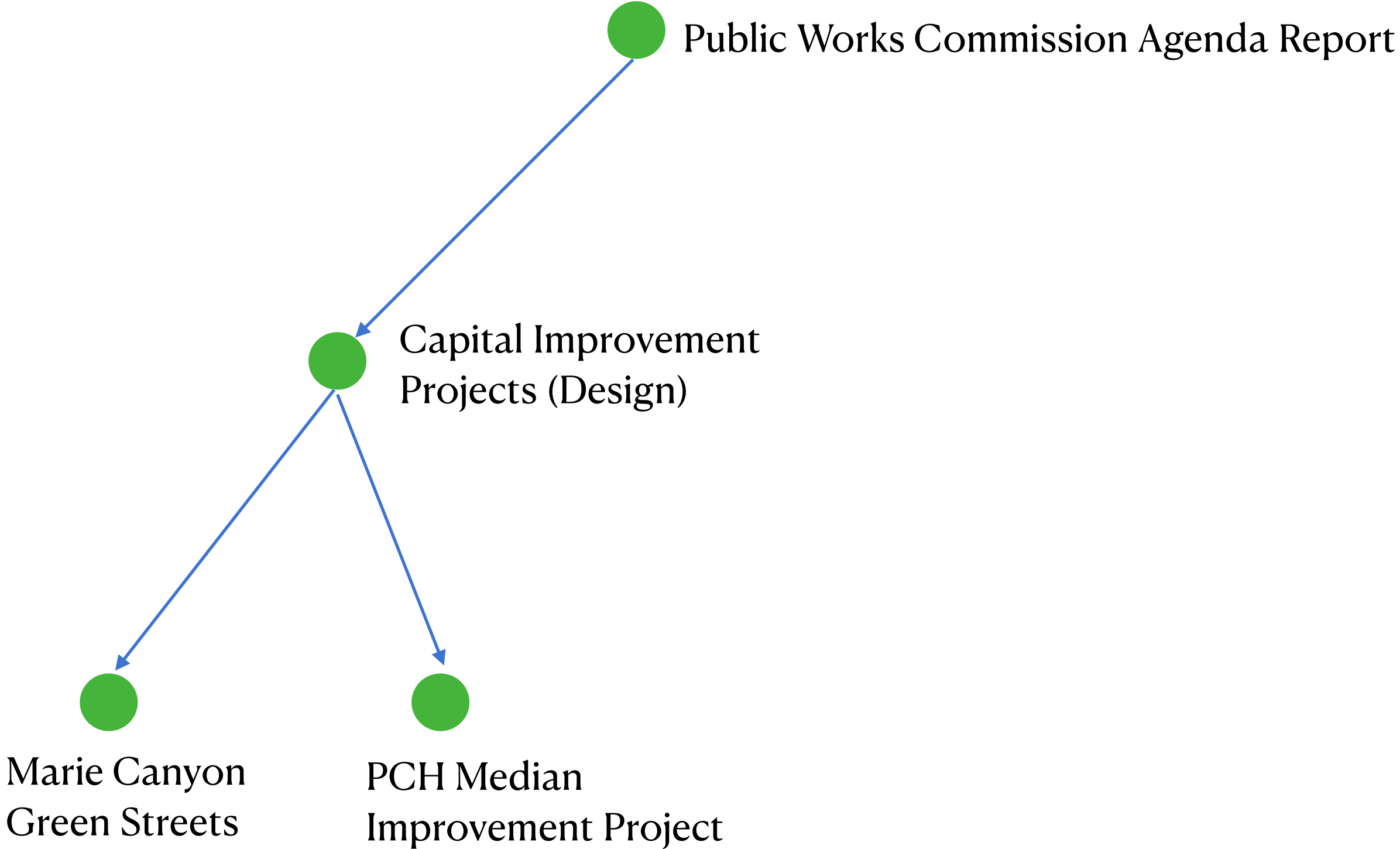
- > Updates:
 - A hydrology report was prepared and will be used to size the pre-manufactured biofilters. City staff reviewed multiple biofilter manufacturers for filters that will work in the proposed project area. The final design is complete, and the project is advertised for construction bids.
- > 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.

Capital Improvement
Projects (Design)

Marie Canyon Green
Streets



Structure In an UnStructured World?

Unstructured Documents are often **semantically structured!**

Jump a few pages...

2021 Annual Street Maintenance

- **Updates:** This project included resurfacing Malibu Road, Broad Beach Road, Latigo Canyon Road, Corral Canyon Road, Webb Way, Rambla Pacifico Street and Vista Pacifica with a slurry seal treatment and adding speed humps to Birdview Avenue. This project was identified in the City's Pavement Management Plan. This project was accepted by the Council at the January 24, 2022 meeting.

Disaster Projects (Design)

Broad Beach Road Water Quality Infrastructure Repairs (CalJPIA Project)

- **Updates:** The project consultant prepared the specifications for the project. The City received bids on April 7, 2022. The agreement is scheduled to go to Council on June 13, 2022 and construction will begin shortly after.
- **Project Schedule:**
 - Complete Design: March 2022
 - Advertise: April 2022
 - Begin Construction: Summer 2022

Latigo Canyon Road Roadway/Retaining Wall Improvements (FEMA Project)

- **Updates:**
 - Staff finalized the design of this project.
 - Staff is also working with FEMA/CalOES to substitute the existing timber with non-combustible materials and include the replacement of guardrails within the project limits.
- **Project Schedule:**
 - Complete Design: April 2022
 - Advertise: Summer 2022
 - Begin Construction: Summer 2022

Trancas Canyon Park Planting and Irrigation Repairs (CalJPIA/FEMA Project)

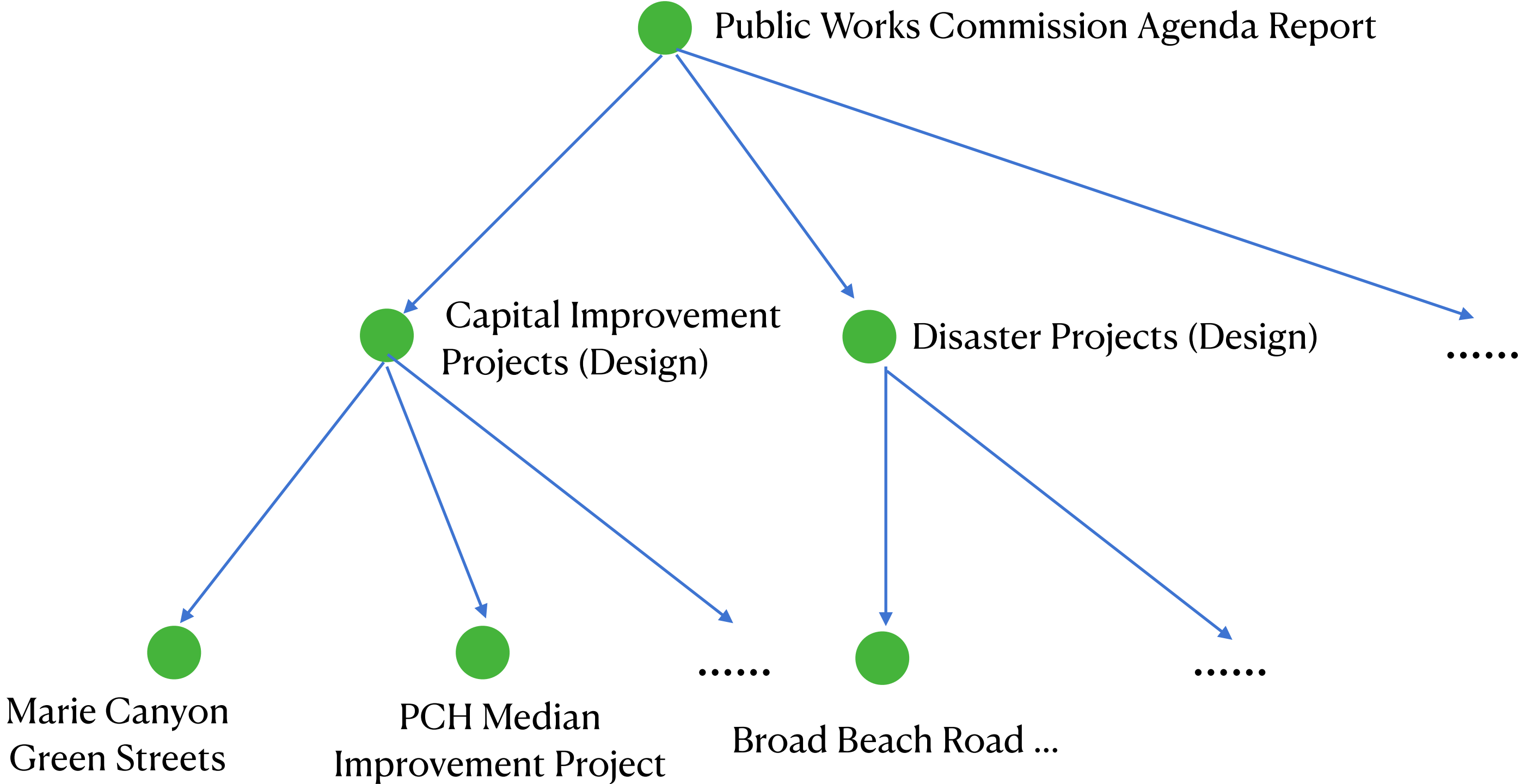
- **Updates:**
 - The project consultant has started the design of this project.
- **Project Schedule:**
 - Complete Design: Spring 2022
 - Begin Construction: Summer 2022

Trancas Canyon Park Slope Stabilization Project (CalJPIA Project)

- **Updates:**
 - The project consultant has started the design of this project.
- **Project Schedule:**
 - Complete Design: Spring 2022
 - Begin Construction: Summer 2022

Disaster Projects (Design)

Broad Beach Road...



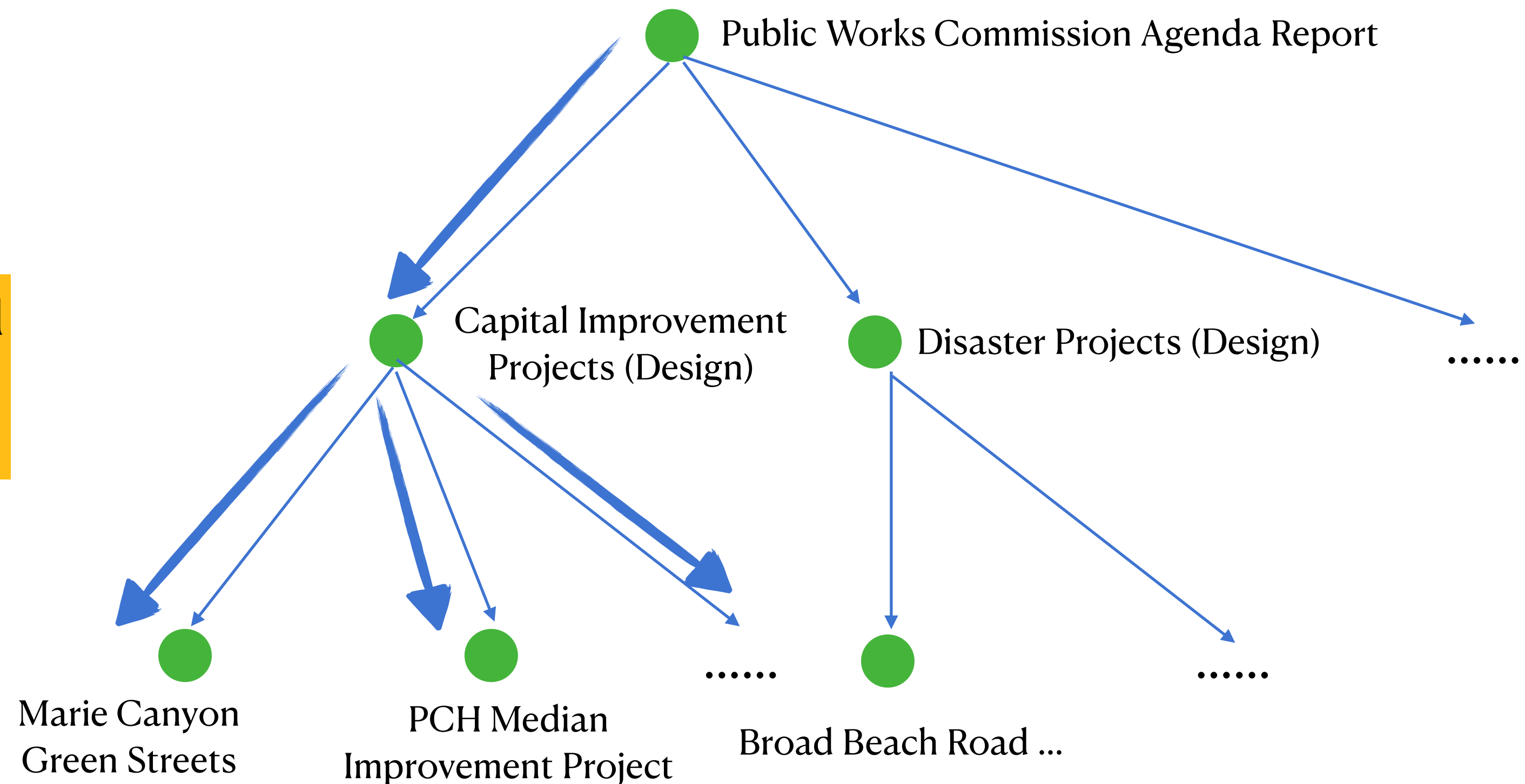
Why Is Semantic Structure Important?

Journalist: give me the number of projects related to **Capital Improvement** starting after 2021.



Imagine how a journalist would read document to figure out the answer?
Akin to “table of contents”

Human-centered Approach



	LLM (GPT-4)	RAG (GPT-4)	ZenDB (GPT-4)	ZenDB (GPT-3.5)
Accuracy	41%	3%	78%	63%
Cost	12.2\$	0.4\$	0.7\$	0.007\$

Templatized Documents

While unstructured documents vary considerably in format,
a significant portion are created by using templates

Living Healthy

#chi4good, CHI 2016, San Jose, CA, USA

Beyond Abandonment to Next Steps: Understanding and Designing for Life after Personal Informatics Tool Use

Daniel A. Epstein¹, Monica Caraway², Chuck Johnston²,
An Ping², James Fogarty¹, Sean A. Munson²

¹Computer Science & Engineering, ²Human Centered Design & Engineering
DUB Group, University of Washington

{depstein, jfogarty}@cs.washington.edu, {mcaraway, chuck2, anping, smunson}@uw.edu

ABSTRACT

Recent research examines how and why people abandon self-tracking tools. We extend this work with new insights drawn from people reflecting on their experiences after they stop tracking, examining how designs continue to influence people even after abandonment. We further contrast prior work considering abandonment of health and wellness tracking tools with an exploration of why people abandon financial and location tracking tools, and we connect our findings to models of personal informatics. Surveying 193 people and interviewing 12 people, we identify six reasons why people stop tracking and five perspectives on life after tracking. We discuss these results and opportunities for design to consider life after self-tracking.

Author Keywords

Personal informatics; self-tracking; abandonment.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI).

INTRODUCTION

Personal informatics is defined as the process of collecting and reflecting on personal information [12], and is now a common practice in the lives of many people [7]. However, people over time come to temporarily lapse or permanently discontinue self-tracking [4,5,6,11]. We study abandonment of self-tracking tools to gain insight into how to design tools that: (1) better align with tracking objectives and practices, and (2) support better abandonment experiences.

This paper extends current understanding of abandonment with insights drawn from people reflecting on their experiences after they stopped self-tracking. As part of this, we examine how designs can continue to influence people even after abandonment. We extend recent work examining self-tracking technology abandonment in health and wellness [4,5,11] by contrasting it with abandonment in other self-tracking domains, specifically finance and location. We

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI'16, May 07-12, 2016, San Jose, CA, USA
© 2016 ACM. ISBN 978-1-4503-3362-7/16/05...\$15.00
http://dx.doi.org/10.1145/2858036.2858045

frame these findings in models of how people use personal informatics tools [6,12], and we identify and discuss how self-tracking barriers lead to abandonment.

We survey 193 people who formerly tracked their physical activity, finances, or location, conduct 12 interviews, and distill themes from this qualitative data. We extend prior work by identifying six reasons people stop tracking and five perspectives on life after tracking among the three studied domains. Our results contribute to a growing understanding of self-tracking abandonment, and surface opportunities for design to consider life after tracking.

BACKGROUND AND RELATED WORK

To characterize how people use self-tracking tools, Li et al. introduce a five-stage model of personal informatics, which emphasizes barriers to tracking toward a goal of reflection and presumed action [12]. This model has been modified and expanded, noting people can reflect on data [3] and ultimately change habits [16] in the midst of tracking. Epstein et al. characterize challenges in lived informatics [14], developing a model of tool use in everyday life that surfaces lapsing and stopping as major components [6].

Avoiding or discontinuing the use of technology is common practice. Baumer et al. enumerate motivations for not using Facebook, including concerns for data use and privacy as well as avoiding addiction [2]. In the domains of physical activity and health and wellness more broadly, recent work has explored reasons people stop using self-tracking tools. Schwanda et al. interview people using Wii Fit, finding they begin other exercise activities and abandon the technology, regarding the abandonment as a success [15]. Clawson et al. extend this “happy abandonment,” suggesting designs should support people who no longer feel the need to track [4]. Tools sometimes satisfy people’s curiosity about their habits, rendering tracking no longer important [6,11]. People find tools frustrating or time-consuming, ultimately not worth the time investment [5,11]. We extend and contrast such findings in the domains of finance and location, building a broader understanding of abandoning tracking.

DATA COLLECTION AND ANALYSIS

We conducted a series of surveys using Amazon Mechanical Turk with people from the United States who had completed at least 1,000 HITs with a 95% acceptance rate. To identify people who had previously tracked, participants completed a 2-minute screener survey (\$0.50 compensation). Of 640

Two papers from the same conferences follow the same template:


- Same visual patterns on headers for all sections...

Templatized Documents

While unstructured documents vary considerably in format, **a significant portion are created by using templates**

Living Health

Beyond Design



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.

Civic documents serving the same purpose from the same local county use the same template

Templatized Documents

While unstructured documents vary considerably in format, a significant portion are created by using templates

Living Health
Beyon Des
{dep
ABSTRACT
Recent resear
self-tracking t
drawn from pr
stop tracking,
people even a
work consid
tracking tools
financial and
findings to m
people and in
why people st
tracking. We
design to cons
Author Keyw
person intro
ACM Classifi
H.S.M. Inform
INTRODUCTI
Personal intro
and reflecting
common pract
people over ti
discontinue se
of self-trackin
that: (1) bette
and (2) suppo
This paper ex
with insights
experiences a
we examine l
even after aba
self-tracking t
[4,5,11] by
self-tracking d
Permission to ma
personal or clas
not made or distri
bear this notice
components of th
Abstracting with
post on servers o
and/or a fee. Requ
CHP16, May 07-
© 2016 ACM. IS
http://dx.doi.org/1

CITY OF M
COMMONWEALTH OF KENTUCKY
ENERGY AND ENVIRONMENT CABINET
DEPARTMENT FOR ENVIRONMENTAL PROTECTION
Division of Water

NOTICE OF VIOLATION

To: KY Utilities Co - Brown Station
ATTN: Mr. Brian Sumner
815 Dix Dam Rd
Harrodsburg, KY 40330

AI Name: KY Utilities Co - Brown Station AI ID: 3148 Activity ID: ENV20140001
Discovery ID: CIV20140002 County: Mercer
Enforcement Case ID:
Date(s) Violation(s) Observed: 03/07/2014

This is to advise that you are in violation of the provisions cited below:

1 Violation Description for Subject Item A100000003148:
Applicability of the KPDES requirements. (1) A KPDES permit shall be required to discharge pollutants from a point source into waters of the Commonwealth. (2) Compliance with the KPDES program requirements shall constitute compliance with the operational permit requirements of 401 KAR 5:005. (3) Failure to obtain a KPDES permit shall not relieve a discharger whose discharge is subject to the KPDES program from complying with the applicable performance standards of the KPDES program, 401 KAR 5:050 through 5:080. [401 KAR 5:055 Section 2]

Description of Non Compliance:
A pipe discharge to the Outfall #3 discharge channel was located by DOW personnel during the 3/7/14 investigation. This discharge was described by facility personnel as an abutment drain from the old ash pond dam. The water from this discharge appeared clear and odorless at the time of the investigation, however there was a red/orange residue at the discharge point from this pipe and within the length of the Outfall #3 discharge channel to Herrington Lake. The facility has a point source discharge not included within the active KPDES permit.

The remedial measure(s), and date(s) to be completed by are as follows:
Cease all unpermitted discharges. Within 30 days of the receipt of this notice, the permittee shall submit a written notification to the undersigned to address the point source discharges. This could include either modifying the current permit to include these point source discharges or modifying the BMP plan. Failure to comply with these remedial measures or repeated violations of this requirement may subject you and your company to an immediate referral to the Division of Enforcement. [401 KAR 5:055 Section 2]

2 Violation Description for Subject Item A100000003148:
The permittee shall report any noncompliance which may endanger health or the environment. Any information shall be provided orally within 24 hours from the time the permittee becomes aware of the circumstances. A written submission shall also be provided within 5 days of the time the permittee becomes aware of the circumstances. The written submission shall contain a description of the noncompliance and its cause; the period of noncompliance, including exact dates and times, and if the noncompliance has not been corrected, the anticipated time it is expected to continue; and steps taken or planned to reduce, eliminate, and prevent reoccurrence of the noncompliance. [401 KAR 5:065 Section 2(1) as in 40 C.F.R. 122.41(l)(6)].

Description of Non Compliance:
A high volume of red/orange water discharging within the Outfall #3 drainage channel and into Herrington Lake with visual evidence of discoloration to Waters of the Commonwealth, was observed by cabinet personnel on 2/20/14. The facility has failed to report non-compliance as required by 401 KAR 5:065 Section 2(1).

RECOM upcomi
DISCU: Fiscal Y
Capital
Marie C
PCH M

Agenda Item # 4.A.

Notice of Violation documents generated from the same department for the same purpose follow the same template

We have similar observations in another 16 datasets from 6 domains

Overview of ZenDB

ZenDB: A DB system for unstructured documents that leverages semantic structure

How to design user-friendly interfaces?

How to ensure user trust in query answer?

DBA/end users write queries

Provenance

Document Ingestion

Query Specification

Query Execution

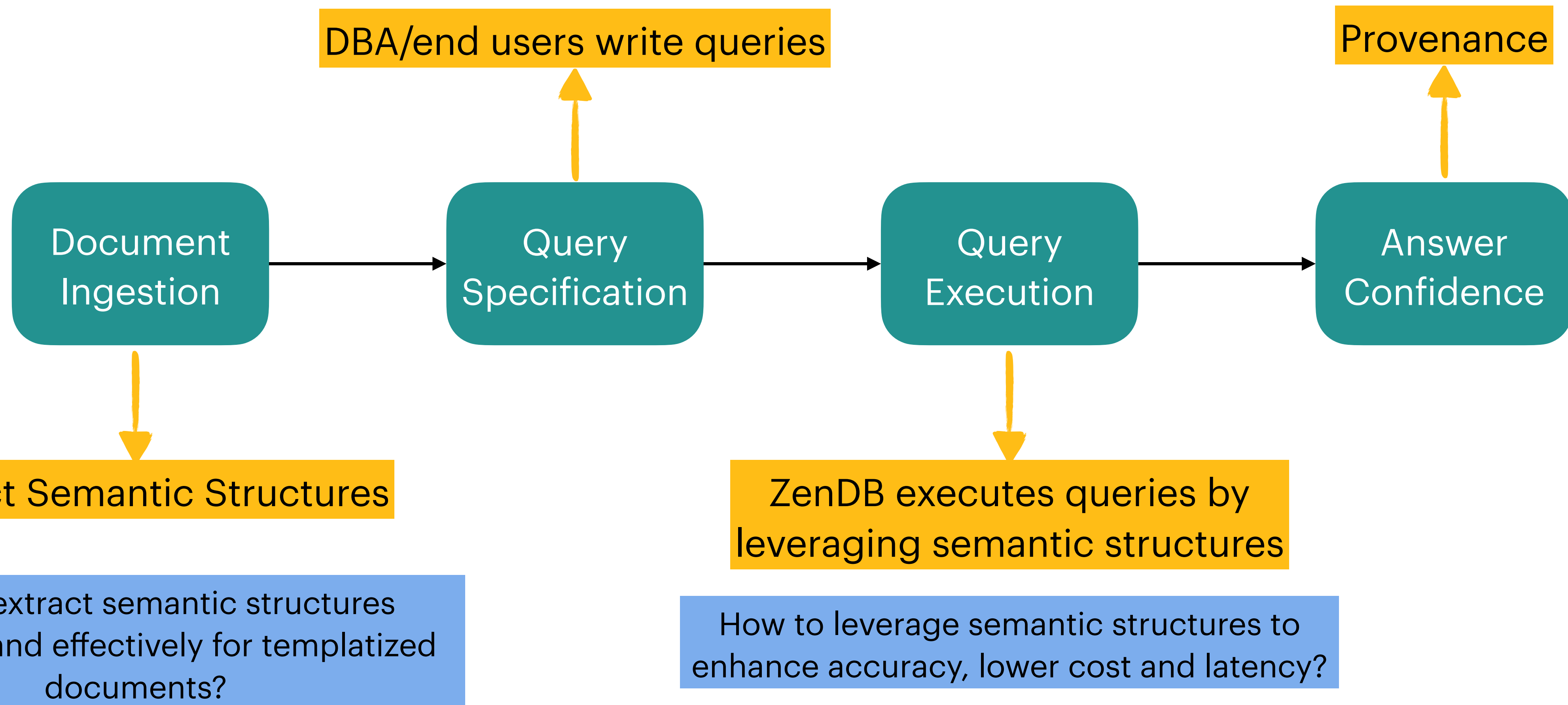
Answer Confidence

Extract Semantic Structures

ZenDB executes queries by leveraging semantic structures

How to extract semantic structures efficiently and effectively for templated documents?


How to leverage semantic structures to enhance accuracy, lower cost and latency?



How Do We Extract Semantic Structure?

- **Visual Pattern:** font size, font name, font type, all capital or not, start from a number or not, begin with alpha or not, centered or not...
- **Oracle:** LLM or human: header or not?

R



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

**Public Works Commission
Agenda Report**

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.

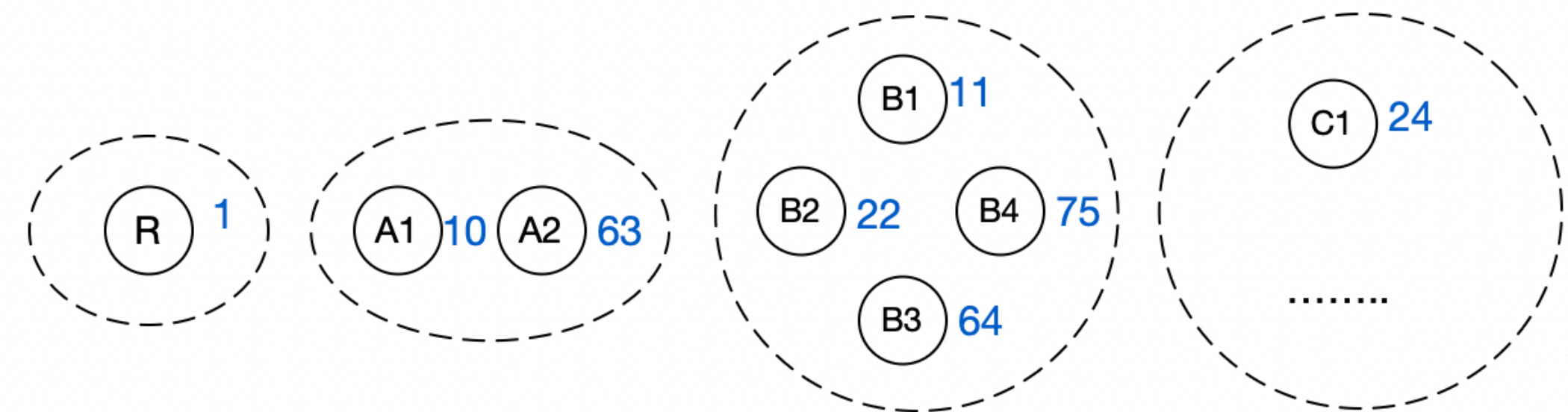
A1

B1

B2

C1

Step 1: Clustering based on visual patterns




How to Extract Semantic Structures?

- **Visual Pattern:** font size, font name, font type, all capital or not, start from a number or not, begin with alpha or not, centered or not...
- **Oracle:** LLM or human: header or not?

Step 2: Remove non-header clusters by asking an LLM oracle

R



**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:

A1

Capital Improvement Projects (Design)

B1

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

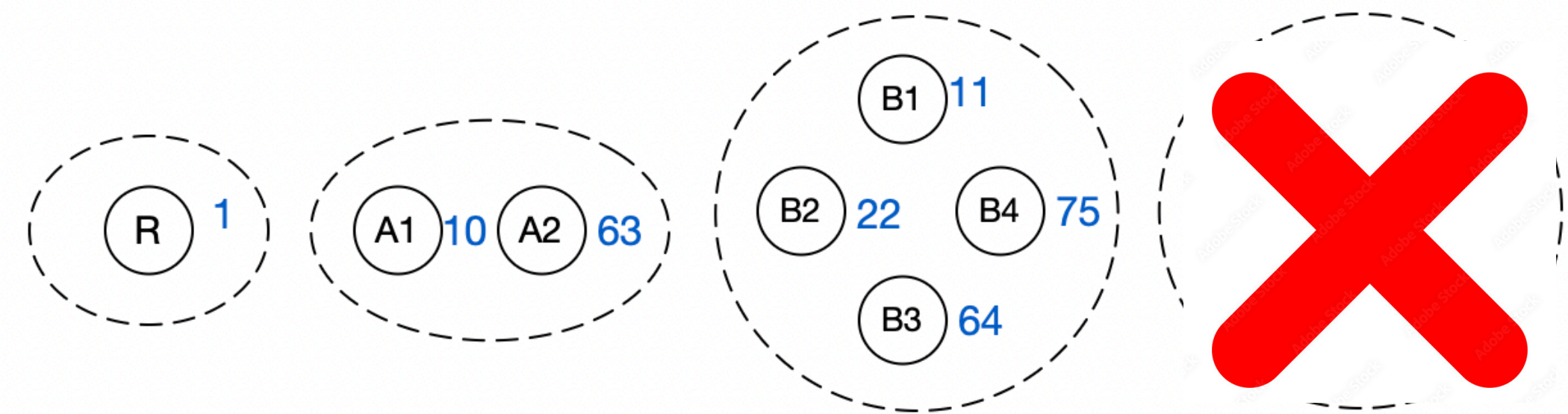
B2

PCH Median Improvements Project

C1

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.



How to Extract Semantic Structures?

- **Visual Pattern:** font size, font name, font type, all capital or not, start from a number or not, begin with alpha or not, centered or not...
- **Oracle:** LLM or human: header or not?

R

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

Public Works Commission Agenda Report

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.

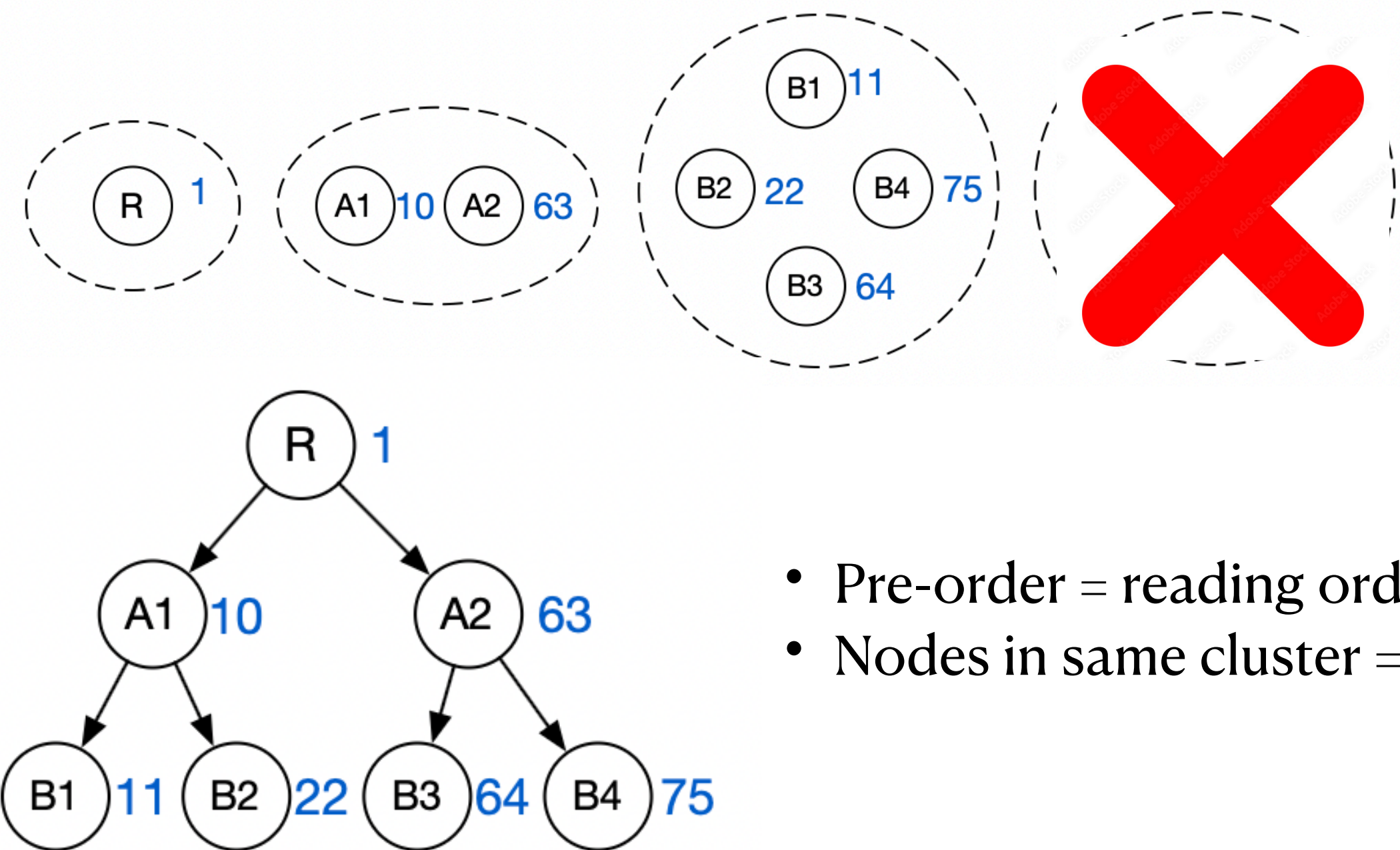
A1

B1

B2

C1

Step 3: Tree Construction



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

Documents sharing templates - by **non-LLM** visual patterns matching

User Interfaces? - Extending SQL

How do we let users write queries in a user-friendly manner?

Journalist: give me the number of projects related to **Capital Improvement** starting after **2021**.

Define Tables

```
CREATE DTABLE TableName AS (description)
```

```
CREATE DTABLE Projects AS  
(‘This table contains a set of projects related with  
Public work commission.’)
```

```
SELECT COUNT(Projects.name)  
FROM Projects  
WHERE Projects.type = ‘Capital Improvement’  
AND Projects.start_date > ‘2021’
```

Define Attributes

```
CREATE ATTRIBUTE ATTRName on TableName AS  
(description, type)
```

```
CREATE ATTRIBUTE Name on Projects AS  
(‘Name of project’, text)
```

```
CREATE ATTRIBUTE Type on Projects AS  
(‘Type of project’, text)
```

```
CREATE ATTRIBUTE start_date on Projects AS  
(‘Start date of project’, date)
```

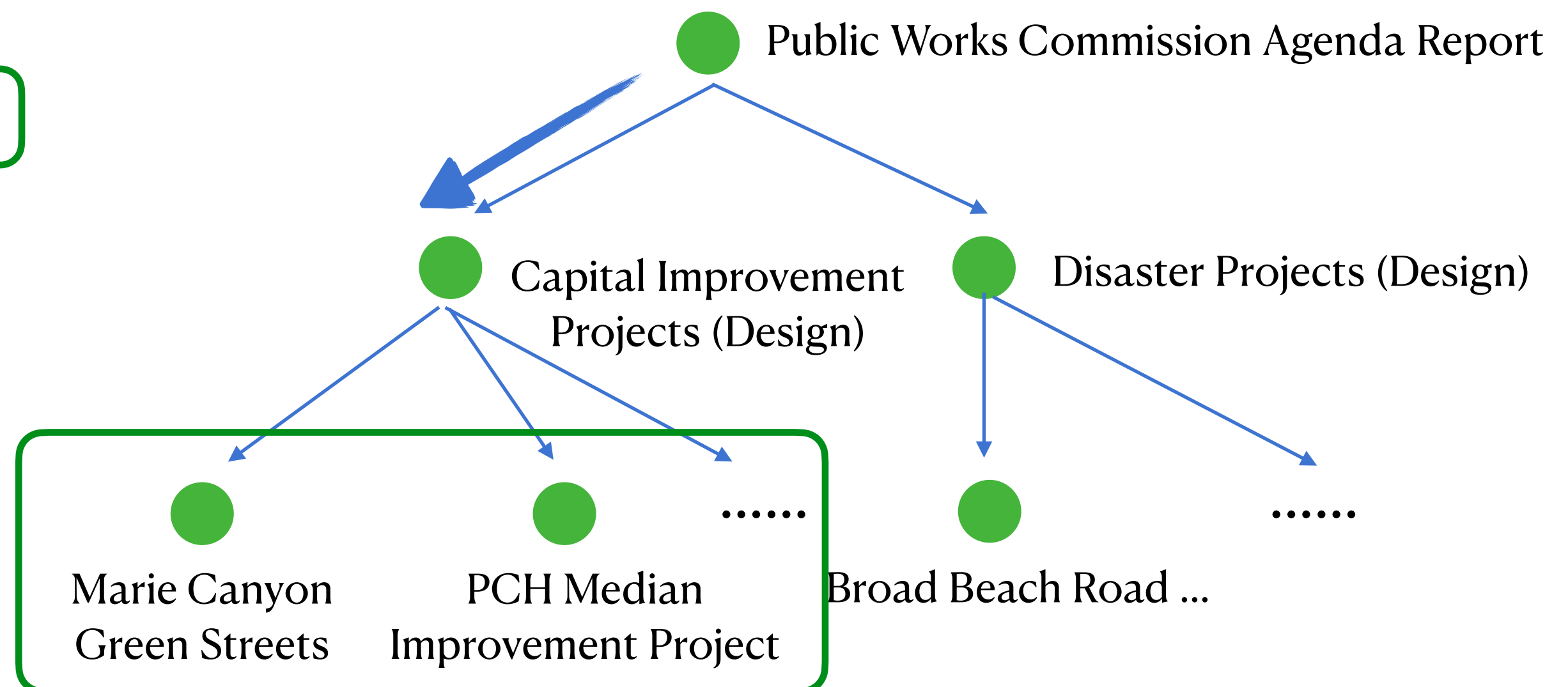
Query Engine

Physical Query Plan: tree-search based on semantic tree

```
SELECT COUNT(Projects.name)
FROM Projects
WHERE Projects.type = 'Capital Improvement'
AND Projects.start_date > '2021'
```

Sketch for each node:

- Name of current node and ancestors
- Summary
- Top-1 sentence similar to queried condition



Can you trust the answers returned by LLM?

Journalist: give me the number of projects related with **Capital Improvement** and **starting after 2021**.

```
SELECT COUNT(Projects.name)
FROM Projects
WHERE Projects.type = 'Capital Improvement'
AND Projects.start_date > '2021'
```

If "Marie Canyon Green Design" is part of the answer,
Can you trust it?

Provenance

- **What is size of provenance?**
 - Too short: no context
 - Too long: lots of human effort
- **What are the right information in provenance?**
 - To verify all filters? Projected attributes?
- **What if the aggregation query?**
 - Aggregate over a large number of tuples?

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

Take-aways from Experimental Study

Datasets: 1) Civic Agenda Reports, 2) Scientific Papers; 3) Notice of Violations.

- **ZenDB VS LLM** (all GPT-4-32k)
 - Up to **+29%** precision, **+31%** recall, **30x** saving of costs, **4x** latency saving
- **ZenDB VS RAG** (all GPT-4-32k)
 - 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 **4.5k SQL Queries** on average on a single document, with usable quality (**-7%** precision and **-5%** recall VS ZenDB + GPT-4-32k)

Conclusion & Future Work

ZenDB: Our first exploration towards building a data management system for **unstructured** documents that leverages **semantic structure**

- What is next?
 - How can we make semantic structure extraction more **robust** in **noisy** document?
 - How can we formulate the **provenance** to **ensure human trust**, going **beyond SQL** interfaces?
 - Scalability? Extending to millions of documents
 -

ZenDB (long-term goal): Towards an **accurate, efficient** and **cost-effective** data management system for **unstructured** documents that support **ad-hoc advanced** analytics

Thanks!
Q&A