HarvestMatch

Mona Khosla, Tanvi Pabbathi, Siri Nellutla, Hugo Leung

CIS 5500 Final Presentation

May 9, 2025

Presentation overview

- Motivation
- Datasets
- Relation Schema
- Demo
- Queries + Performance
- Challenges

Motivation

Farmers, gardeners, and agricultural planners often lack accessible, region-specific crop guidance that accounts for complex environmental variables.

What crops grow best in my state?

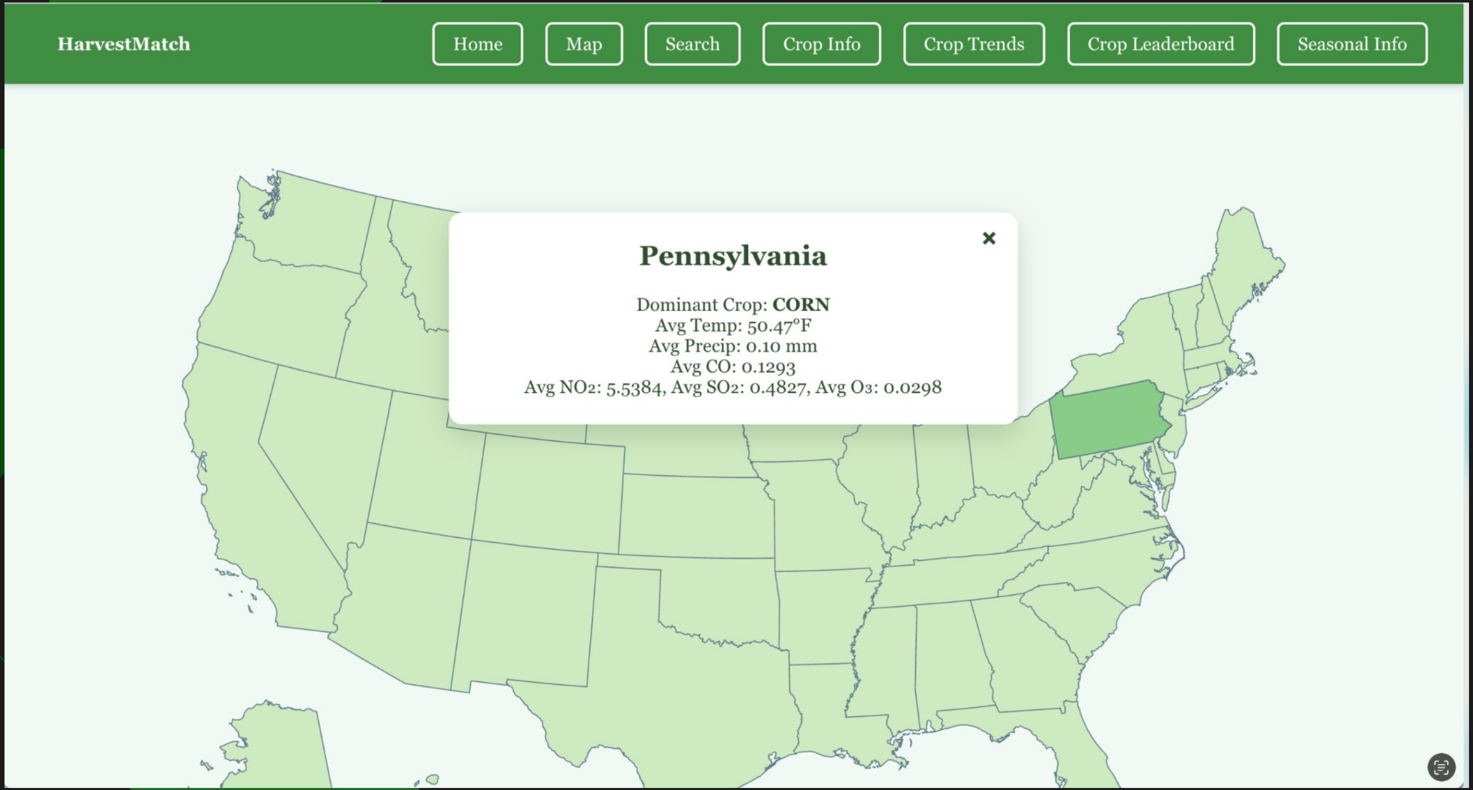
What crop is most resilient to climate change?

Which crops thrive under current pollution, temperature, or precipitation?

When should I plant a specific crop?

Questions you can ask HarvestMatch

Snapshot


HarvestMatch
Home
Map
Search
Crop Info
Crop Trends
Crop Leaderboard
Seasonal Info
Pennsylvania
Dominant Crop: CORN
Avg Temp: 50.47°F
Avg Precip: 0.10 mm
Avg CO: 0.1293
Avg NO2: 5.5384, Avg SO2: 0.4827, Avg O3: 0.0298

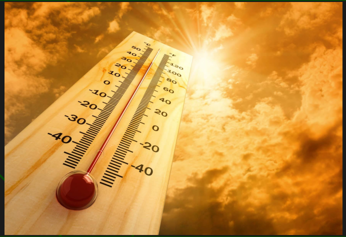
Datasets

Pollution dataset

Tracks daily air quality metrics (O_3, CO, SO_2, NO_2) by state and city
It helps identify how pollution levels affect crop performance and resilience.

Crops dataset

Includes historical crop yields (kg/acre) by state, crop type, and month. This dataset enables comparison of crop productivity across time and regions.



Temperature dataset

Captures monthly average temperatures per state with geographic coordinates. The temperature dataset supports climate-based crop suitability analysis.

Weather dataset

Captures historical weather events across regions in the US. Each event includes location, severity, precipitation, and timing data.

We built a web app that empowers users to explore crop suitability using three interlinked datasets across climate, environment, and agricultural yield.

Pollution Data Pre-processing

- Cleaned missing AQI and mean values
- Mapped pollution readings to seasons
- Dropped columns irrelevant to our application
- Removed null values

Crop Data Pre-processing

- Normalized crop yield units to kg/acre
- Month to Season Conversion for ease of joins on other relations
- Removed null values

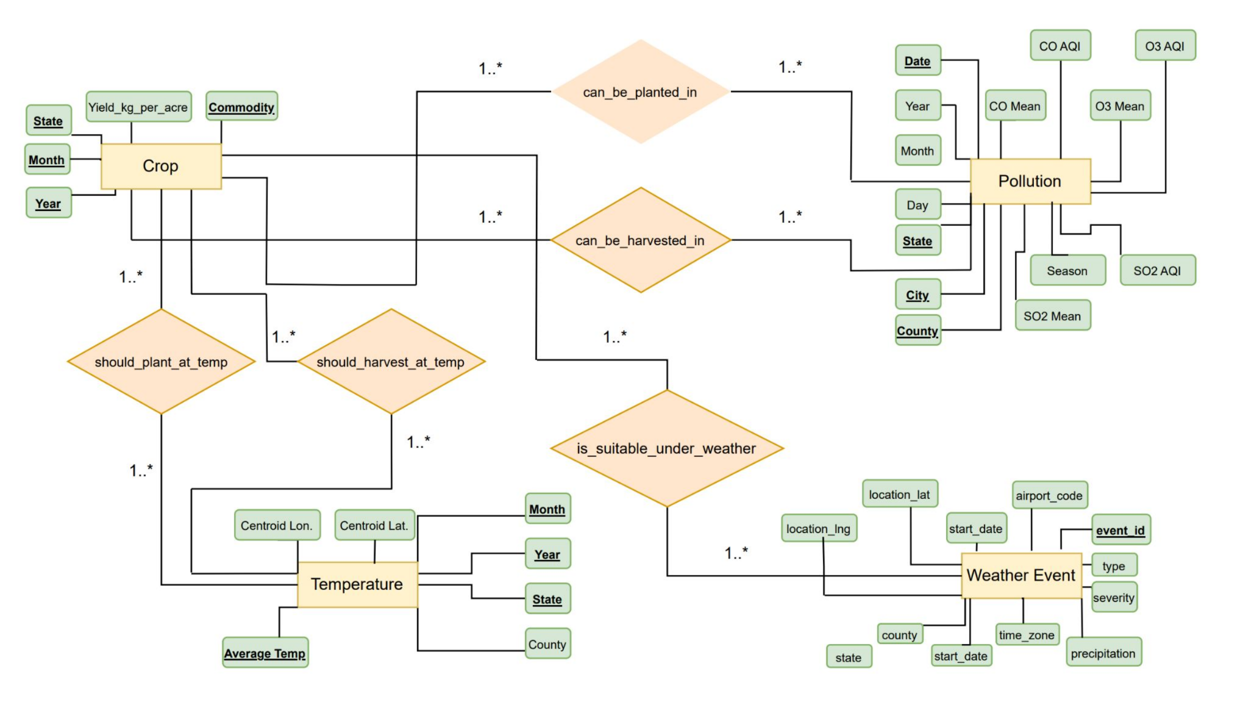
Temperature Data Pre-processing

- Mapped months to their respective seasons in order to join on the variable with other relations
- Removed null values

Weather Data Pre-processing

- Converted state abbreviations to names in order to perform joins on the other relations
- Assigned seasons based on month
- Dropped irrelevant columns
- Removed null values

Entity Relationship Diagram (3NF)



Demo!

Complex Query Example 1: Most to Least Climate Resilient Crops (35s)

```
WITH yearly_precip AS (
  SELECT
   EXTRACT(YEAR FROM start_date)::int AS year, UPPER(state) AS state, AVG(precipitation) AS avg_precip
   FROM weather_events
  WHERE start_date BETWEEN '2016-01-01' AND '2022-12-31'
  GROUP BY EXTRACT(YEAR FROM start_date), UPPER(state)
), crop_env AS (
  SELECT
    c.crop, c.yield_kg_per_acre,  (p."CO Mean" + p."NO2 Mean" + p."SO2 Mean" + p."O3 Mean") AS pollution,  t.average_temp, y.avg_precip
  FROM crop_data c
  JOIN pollution_data p ON c.year = p."Year" AND UPPER(c.state) = UPPER(p."State")
  JOIN temperature_data t ON c.year = t.year AND UPPER(c.state) = UPPER(t.state)
  JOIN yearly_precip y ON c.year = y.year AND UPPER(c.state) = y.state
  WHERE c.year BETWEEN 2016 AND 2022
), classified AS (
  SELECT
    crop, yield_kg_per_acre,
    CASE
      WHEN pollution < 15 OR pollution > 35 THEN 1 ELSE 0
    END +
    CASE
      WHEN average_temp < 15 OR average_temp > 25 THEN 1 ELSE 0
    END +
    CASE
      WHEN avg_precip < 400 OR avg_precip > 900 THEN 1 ELSE 0
    END AS extreme_score
  FROM crop_env
), crop_resilience AS (
  SELECT
    crop, AVG(yield_kg_per_acre) FILTER (WHERE extreme_score >= 2) AS avg_yield_in_extremes
  FROM classified
  GROUP BY crop
  HAVING COUNT(*) FILTER (WHERE extreme_score >= 2) > 1
) SELECT
  crop,
  ROUND(avg_yield_in_extremes::numeric, 2) AS avg_yield_in_extremes
FROM crop_resilience
ORDER BY avg_yield_in_extremes DESC;
```

Complex Query Example 2: Avg Crop Yield Based on Avg Pollution, Precipitation, and Temperature (27s)

```
WITH crop_yearly AS (
 SELECT year,UPPER(state) AS state, AVG(yield_kg_per_acre) AS avg_yield
 FROM crop_data
 WHERE year BETWEEN 2016 AND 2021
 GROUP BY year, UPPER(state)
), pollution_yearly AS (
 SELECT
   "Year" AS year, UPPER("State") AS state, AVG("CO Mean") AS avg_co, AVG("NO2 Mean") AS avg_no2, AVG("SO2 Mean") AS avg_so2, AVG("O3 Mean") AS avg_o3
 FROM pollution_data
 WHERE "Year" BETWEEN 2016 AND 2021
 GROUP BY "Year", UPPER("State")
), precip_yearly AS (
 SELECT
  EXTRACT(YEAR FROM start_date)::int AS year,  UPPER(state) AS state, AVG(precipitation) AS avg_precipitation
 FROM weather_events
 WHERE EXTRACT(YEAR FROM start_date)::int BETWEEN 2016 AND 2021
 GROUP BY EXTRACT(YEAR FROM start_date), UPPER(state)
), temperature_yearly AS (
 SELECT year, UPPER(state) AS state, AVG(average_temp) AS avg_temp
 FROM temperature_data
 WHERE year BETWEEN 2016 AND 2021
 GROUP BY year, UPPER(state)
)SELECT c.year, c.state, ROUND(c.avg_yield::numeric, 2) AS avg_yield, ROUND(p.avg_co::numeric, 4) AS avg_co, ROUND(p.avg_no2::numeric, 4) AS avg_no2,
ROUND(p.avg_so2::numeric, 4) AS avg_so2, ROUND(p.avg_o3::numeric, 4) AS avg_o3, ROUND(w.avg_precipitation::numeric, 2) AS avg_precipitation, ROUND(t.avg_temp::numeric, 2)
AS avg_temp
FROM crop_yearly c
LEFT JOIN pollution_yearly p ON c.year = p.year AND c.state = p.state
LEFT JOIN precip_yearly w ON c.year = w.year AND c.state = w.state
LEFT JOIN temperature_yearly t ON c.year = t.year AND c.state = t.state
ORDER BY c.state, c.year;
```

Performance

Query	Initial Execution Time	Optimized Execution Time
Get historical averages by state and season	12s 86 ms	107 ms
Get avg crop yield based on avg pollution, precipitation, and temperature	26s 593 ms	176 ms
Get best conditions to grow each crop	17s 643ms	170ms
Get a ranking from most to least climate resilient crops	36s 802 ms	321 ms
Best crop to plant based on precipitation	38s 657ms	570ms

Technical Challenges

Massive Datasets

Imported datasets with hundreds of thousands of rows, faced issues with schema design, indexing, and query performance, required data cleaning and normalization before ingestion

Routing to Interactive Map

Managing React Router across multiple interactive components, ensuring state persistence and debugging navigation between map and search pages, Required careful UI-state and query coordination.

Query Optimization

Even with optimized views, high join cardinality (~600K rows) from pollution data required aggregation to ~3K rows to achieve sub-1s query speed.

Thank you!

Image Citations

https://www.nrdc.org/stories/air-pollution-everything-you-need-know

http://rodaleinstitute.org/why-organic/organic-farming-practices/crop-rotations/

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.twinkl.nl%2Fteaching-wiki%2Fweather&psig=AOvVaw0ePDdZrXcYW9goqooepBPv&ust=1746884594623000&source=images&cd=vfe&opi=89978449&ved=0CBcQjhxqFwoTCMjX47DClo0DFQAAAAAdAAAAABAE