Data science and engineering for property tax equity

Nicole Jardine & Jean Cochrane
Cook County Assessor's Office | Data Team

COOK COUNTY ASSESSOR'S OFFICE

Hypothetical average Chicago home in 2015:

Average **estimated value:** **$237,000**

Average **property tax bill:** **$3,600**

The New York Times OPINION

Chicago Tribune

PROPUBLICA

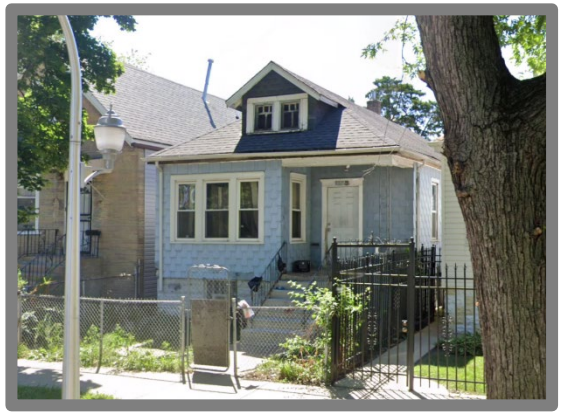
THE UNIVERSITY OF CHICAGO Harris Public Policy

How Lower-Income Americans Get Cheated on Property Taxes

Many homeowners are paying a total of billions of dollars extra because of inequities in assessing property values.

April 3, 2021

Chicago



$100K sale

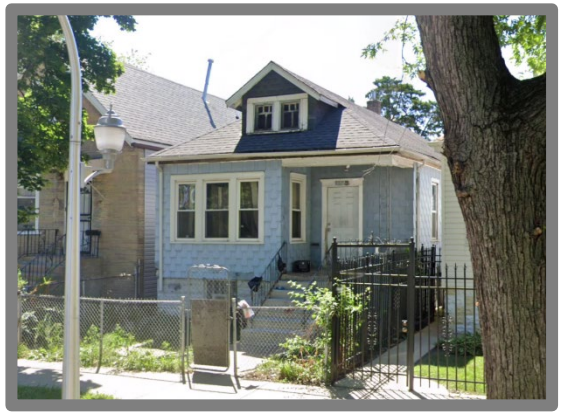
$1M sale

Bottom 10%

Top 10%

Highest price →

Source: **Chris Berry/NYT,** using assessments from prior years



$100K estimate

÷ **$100K** sale

= **1.0 ratio**

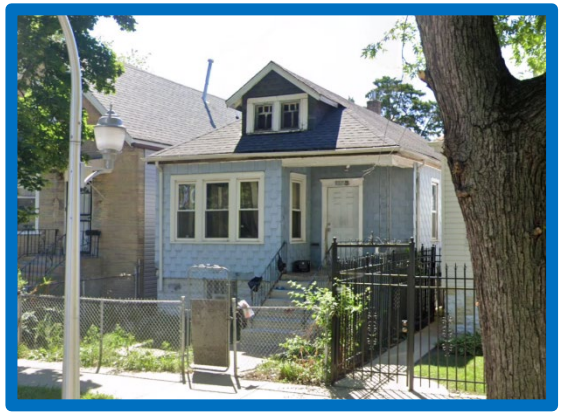
Chicago

$1M estimate

÷ **$1M** sale =

1.0 ratio

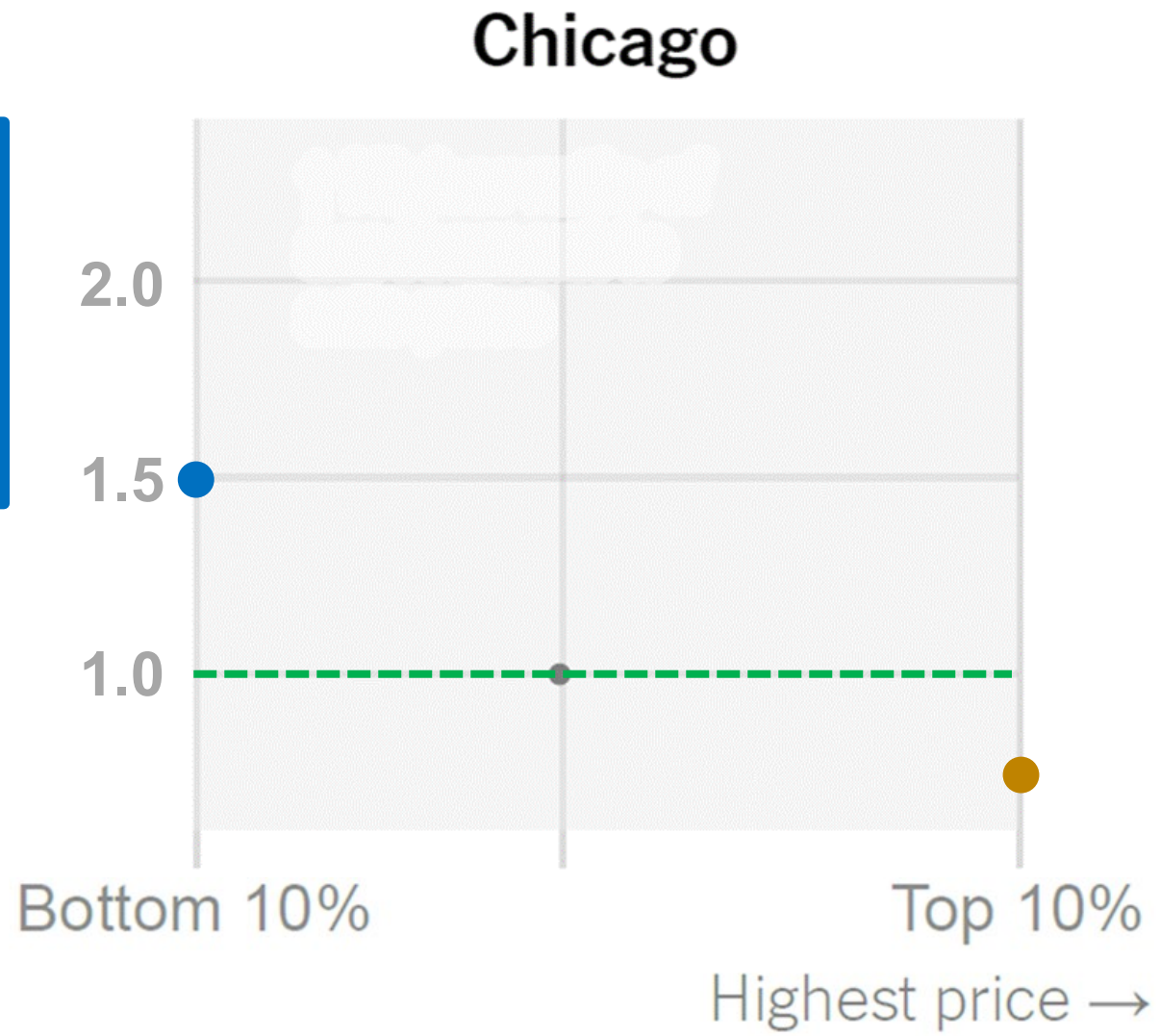
Source: **Chris Berry/NYT,** using assessments from prior years



$150K estimate

÷ **$100K** sale

= **1.5 ratio**

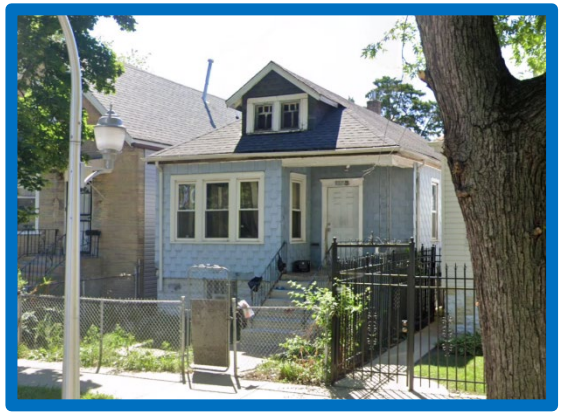


$800K estimate

÷ **$1M** sale

= **0.8 ratio**

Source: **Chris Berry/NYT,** using assessments from prior years



$150K estimate

÷ **$100K** sale

= **1.5 ratio**

$800K estimate

÷ **$1M** sale

= **0.8 ratio**

Source: **Chris Berry/NYT,** using assessments from prior years

Data Science & Analytics: better modeling practices

1. *Better feature engineering*
2. *Better training data*
3. *Better modeling*

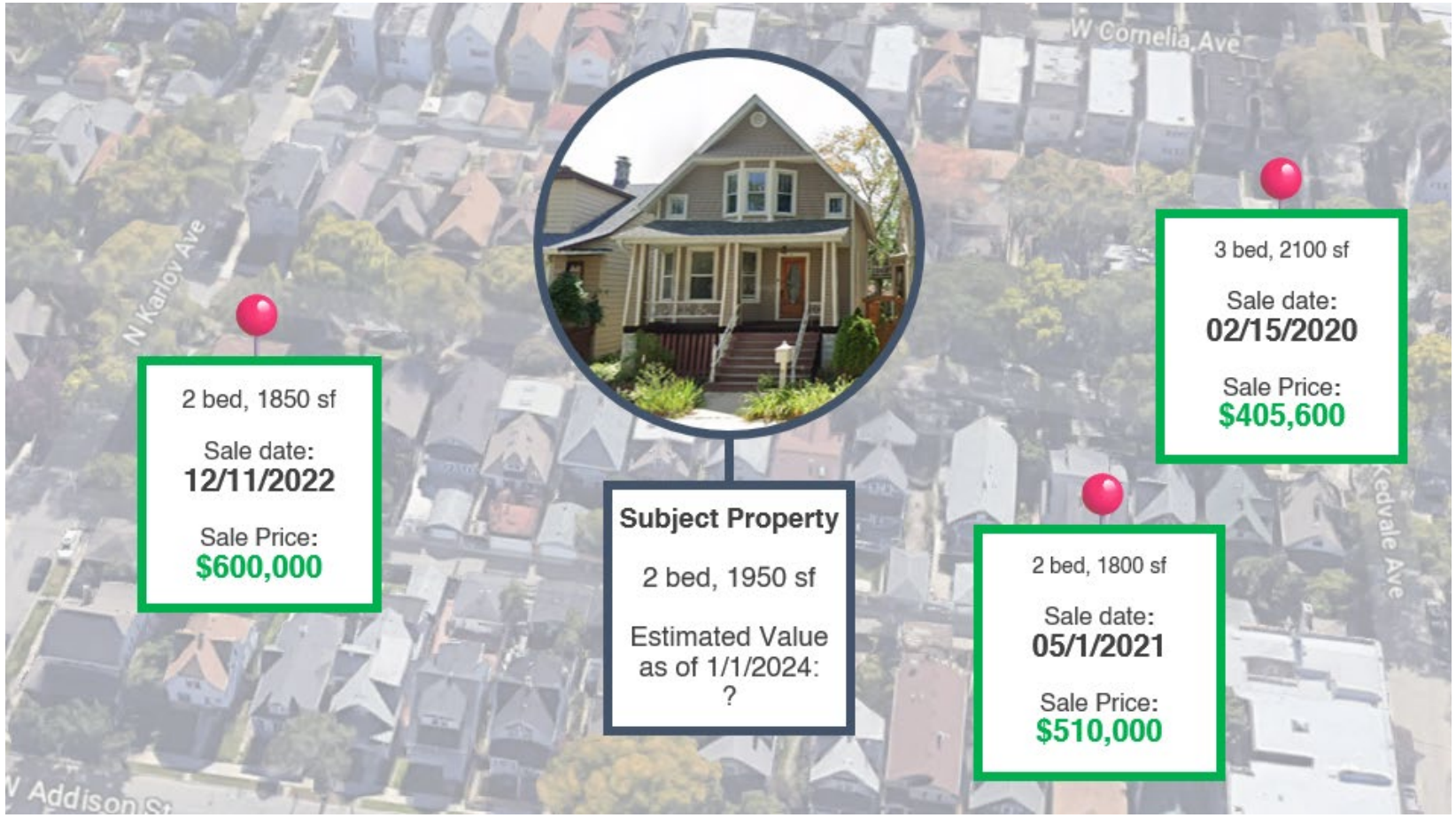
COOK COUNTY ASSESSOR'S OFFICE

Assessing property values 101

The job of the Assessor's Office is *mass* appraisal:

We estimate the property's current market value (as of January 1), every year.

What drives real estate value?



1. Better feature engineering: Location

Then

system of record

Geospatial features

- Assessor neighborhood

Now

system of record + feature engineering

Geospatial features

- Assessor neighborhood
- ✓ Flood risk (First Street)
- ✓ Floodplain (FEMA)
- ✓ Airport noise
- ✓ Proximity to amenities
- ✓ Proximity to disamenities
- ✓ School
- ✓ School ratings
- ✓ Walkability

1. Better feature engineering: Location

Mainframe + AS/400 +

Feature	Data Source
Tax rate	Cook County Clerk's Office
Airport noise	Noise monitoring stations via the Chicago Department of Aviation
Road proximity	Buffering OpenStreetMap motorway, trunk, and primary roads
Flood risk and direction	First Street flood data
All Census features	ACS 5-year estimates for each respective year
Elementary school district or attendance boundary	Cook County school district boundaries and CPS attendance boundaries
High school district or attendance boundary	Cook County high school district boundaries and CPS high school attendance boundaries
Walkability	The Chicago Metropolitan Agency for Planning's ON TO 2050 Walkability Scores
Subdivision, unincorporated areas, SSAs, etc.	Cook County GIS
PUMA Housing Index	DePaul Institute for Housing Studies
School Ratings	GreatSchools.org, aggregated to the district level
Distance to CTA, PACE, Metra	Each agency's respective GTFS feed, which contains the location of stops and lines

Now

system of record + feature engineering

Geospatial features

- Assessor neighborhood
- ✓ Flood risk (First Street)
- ✓ Floodplain (FEMA)
- ✓ Airport noise
- ✓ Proximity to amenities
- ✓ Proximity to disamenities
- ✓ School
- ✓ School ratings
- ✓ Walkability

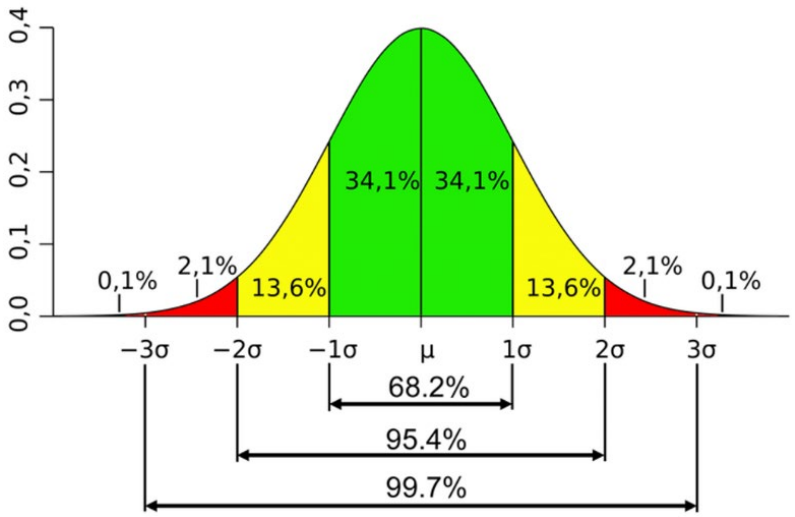
2. Better training data

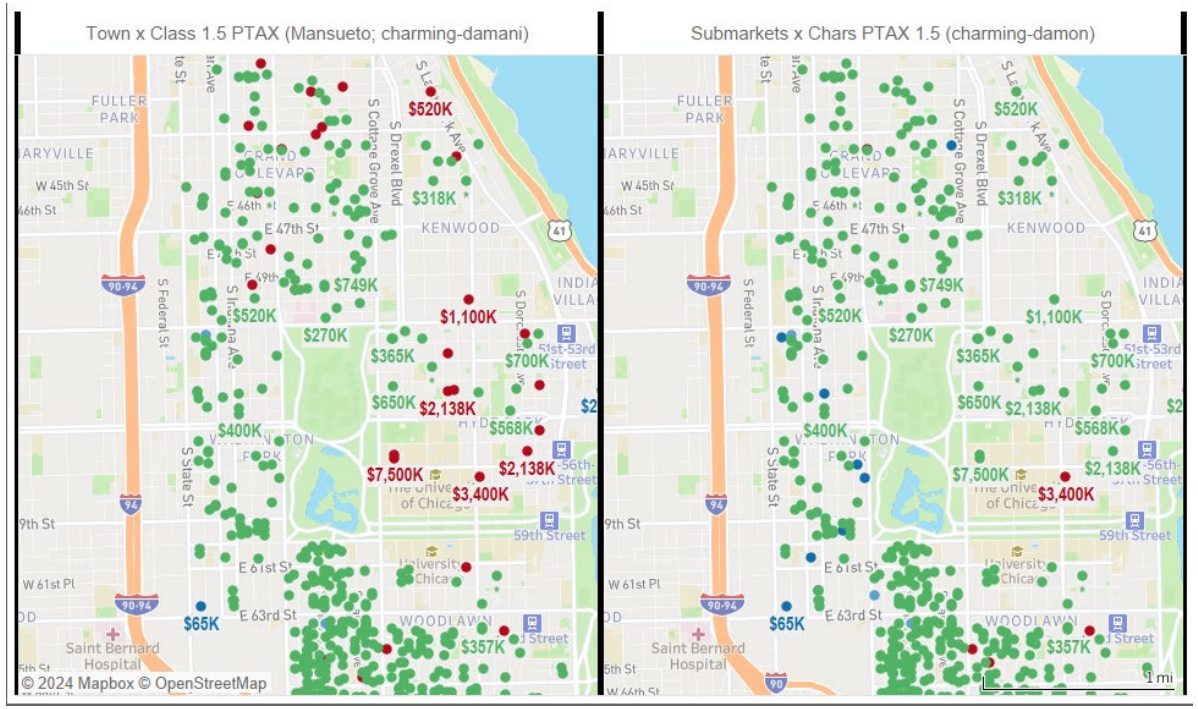
@ccao-data/**model-sales-val**

Problem: some sales are **outliers** (e.g, a home that sells for 50% higher than similar homes.)

Why it matters: Garbage in, garbage out.

Solution: build a pipeline to identify and exclude statistical outlier sales (~7% of sample).





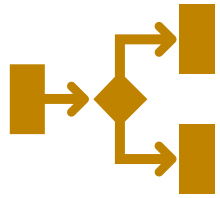
3. Better modeling

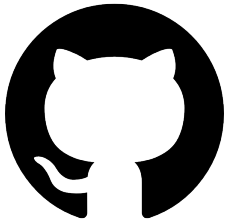
Then

38 linear models

closed-source

Now

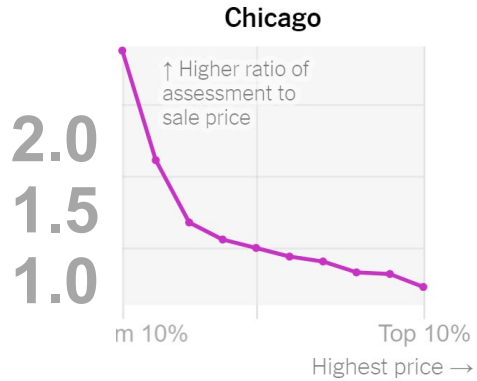




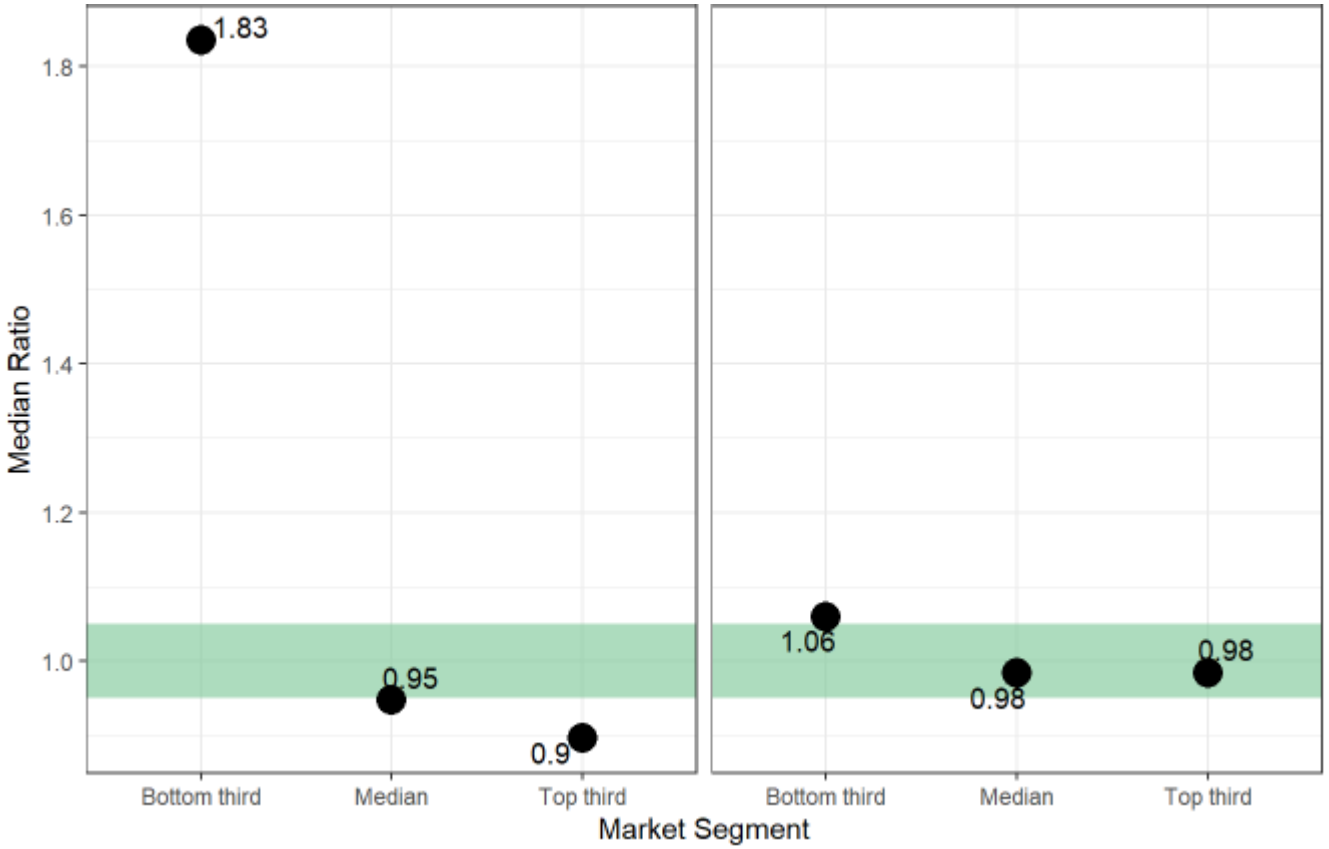
1 tree-based (LightGBM) model

open-source

Results: West Chicago



Then** (2018)* ***Now (2021)



Are we done? Not yet…

DESCRIPTION	Two or more story residence, over 62 years, up to 2,200 sq.ft. 😱
RESIDENCE TYPE	Two Story
USE	Single Family
APARTMENTS	0
EXTERIOR CONSTRUCTION	Masonry
FULL BATHS	2
HALF BATHS	0
BASEMENT[1]	Full
ATTIC	None
CENTRAL AIR	Yes
NUMBER OF FIREPLACES	0
GARAGE SIZE/TYPE[2]	None
AGE	130
BUILDING SQUARE FOOTAGE	2,343 😱

Data Science & Analytics: better modeling practices

- better data science → better public service
- code is policy
- we can institutionalize reform

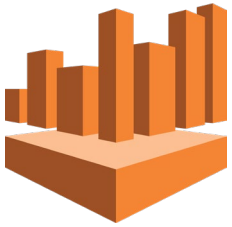
COOK COUNTY ASSESSOR'S OFFICE

Data Science & Analytics:
better modeling practices

Data & Software Engineering:
better infrastructure, cleaner data

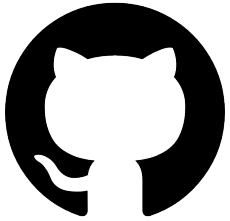
COOK COUNTY
ASSESSOR'S
OFFICE

Better infrastructure



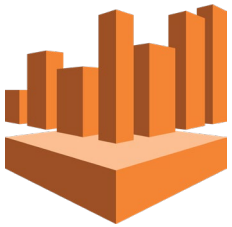
AWS Athena Serverless columnar data storage and retrieval

dbt Data transformation, testing, and documentation



GitHub Actions Code automation

Better infrastructure



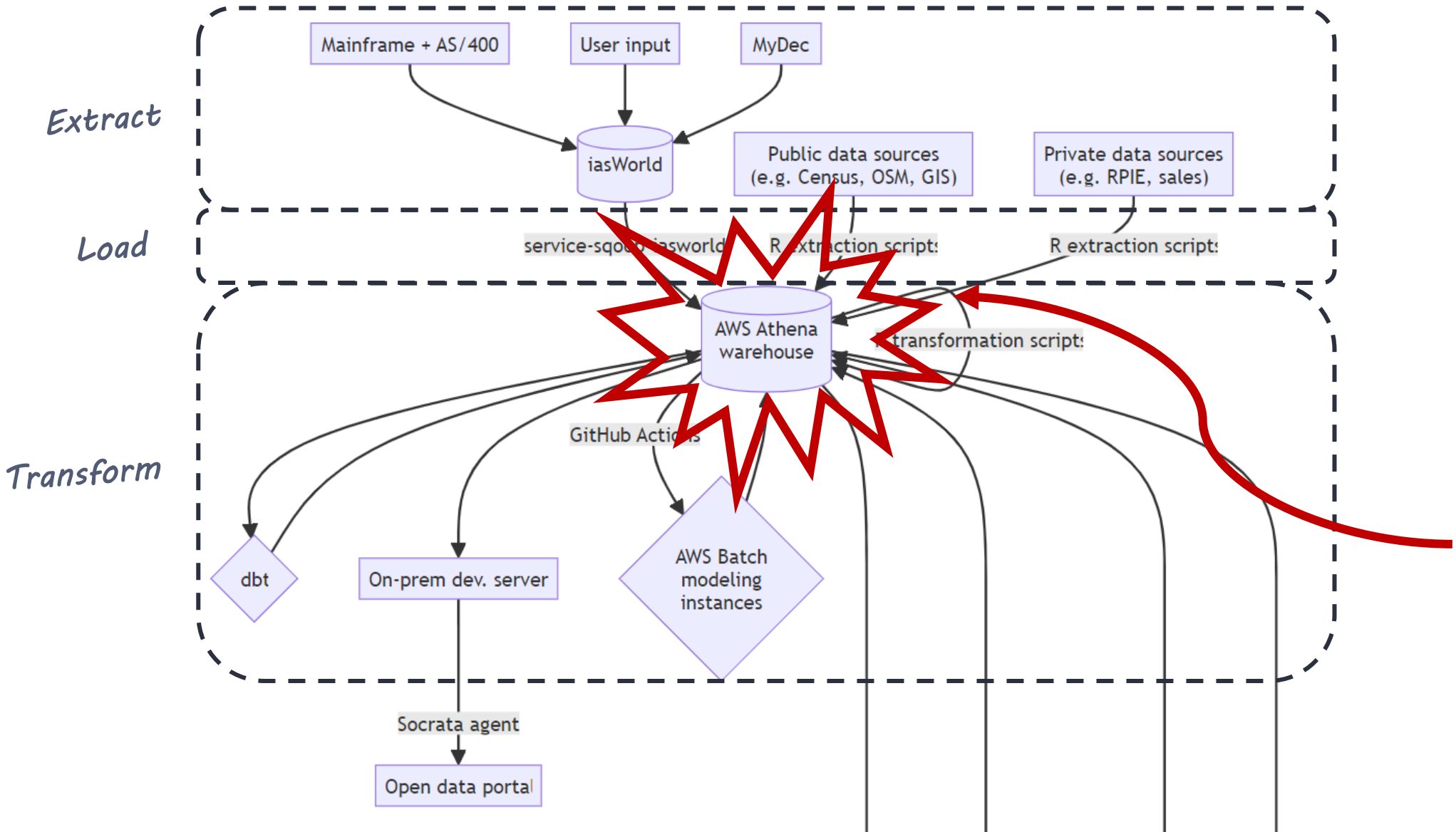
AWS Athena Serverless columnar data storage and retrieval

dbt Data transformation, testing, and documentation

GitHub Actions Code automation

Better infrastructure

Athena



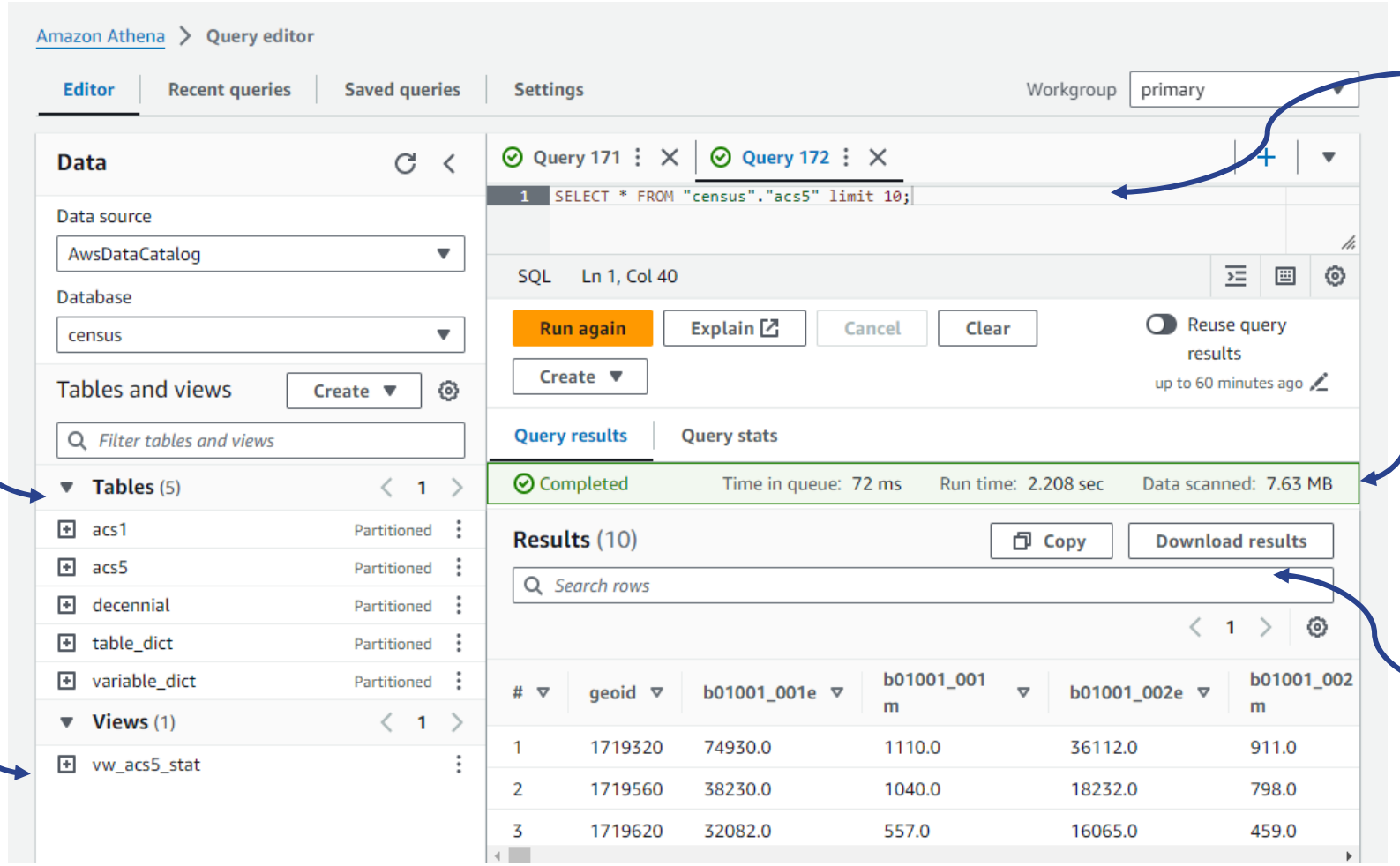
Better infrastructure

Athena

Athena: Serverless columnar database on AWS

Store tables as parquet files in S3 (198 total)

Store views as SQL queries (46 total)



Query data directly from S3

Pay by amount of data scanned

Access results through UI, Python/R SDKs, or Spark jobs

Taxpayer benefit

Athena

⏰ Saves **dev time**: No database management

📈 Saves **uptime**: No maintenance windows

💸 Saves **money**: Cheaper than managed alternatives

Description	Usage Quantity	Amount in USD
Simple Storage Service		**USD 76.85**
US East (N. Virginia)		USD 76.85
Amazon Simple Storage Service Requests-Tier1		USD 18.57
Amazon Simple Storage Service Requests-Tier2		USD 18.67
Amazon Simple Storage Service TimedStorage-ByteHrs		USD 39.61
Athena		**USD 104.11**
US East (N. Virginia)		USD 104.11
Amazon Athena Spark_Execution		USD 43.35
0.35 USD per DPU-Hour for CodeExecutionInDPUHours:Spark_Execution	123.869 DPU-Hour	USD 43.35
Amazon Athena USE1-DataScannedInTB		USD 60.76
5.00 USD per Terabytes for DataScannedInTB in US East (N. Virginia)	12.152 Terabytes	USD 60.76

50m requests

1.8 TB data stored

Better infrastructure

AWS Athena Serverless columnar data storage and retrieval

dbt Data transformation, testing, and documentation

GitHub Actions Code automation

Better infrastructure

dbt

dbt = **D**ata **B**uild **T**ool

Open source!

@dbt-labs/dbt-core

```sql
-- CTAS to create a table of distance to the nearest CTA stop for each PIN
{{
    config(
        materialized='table',
        partitioned_by=['year'],
        bucketed_by=['pin10'],
        bucket_count=1
    )
}}

WITH cta_stop AS (  -- noqa: ST03
    SELECT *
    FROM {{ source('spatial', 'transit_stop') }}
    WHERE agency = 'cta'
        AND route_type = 1
)

SELECT
    pcl.pin10,
    ARBITRARY(xy.stop_id) AS nearest_cta_stop_id,
    ARBITRARY(xy.stop_name) AS nearest_cta_stop_name,
    ARBITRARY(xy.dist_ft) AS nearest_cta_stop_dist_ft,
    ARBITRARY(xy.year) AS nearest_cta_stop_data_year,
    pcl.year
FROM {{ source('spatial', 'parcel') }} AS pcl
INNER JOIN ( {{ dist_to_nearest_geometry('cta_stop') }} ) AS xy
    ON pcl.x_3435 = xy.x_3435
    AND pcl.y_3435 = xy.y_3435
    AND pcl.year = xy.pin_year
GROUP BY pcl.pin10, pcl.year
```

```yaml
- name: proximity.dist_pin_to_cta_stop
  description: '{{ doc("table_dist_pin_to_cta_stop") }}'

- name: proximity.dist_pin_to_golf_course
  description: '{{ doc("table_dist_pin_to_golf_course") }}'

- name: proximity.dist_pin_to_hospital
  description: '{{ doc("table_dist_pin_to_hospital") }}'

- name: proximity.dist_pin_to_lake_michigan
  description: '{{ doc("table_dist_pin_to_lake_michigan") }}'

- name: proximity.dist_pin_to_major_road
  description: '{{ doc("table_dist_pin_to_major_road") }}'
```

```
jecochr@ccao-datals:~/data-architecture$ dbt build --select
proximity.dist_pin_to_cta_stop
```

Better infrastructure

dbt

dbt = **D**ata **B**uild **T**ool

@dbt-labs/dbt-core

We use dbt for...

- 🚀 Continuous integration for data transformations
- 🧪 Data quality testing
- 📖 Data documentation

@ccao-data/data-architecture

Also open source!!

Better infrastructure

dbt

🚀 Continuous integration for data transformations

Before...

😱 Testing done in prod

😬 View changes copy/pasted from GitHub to Athena UI

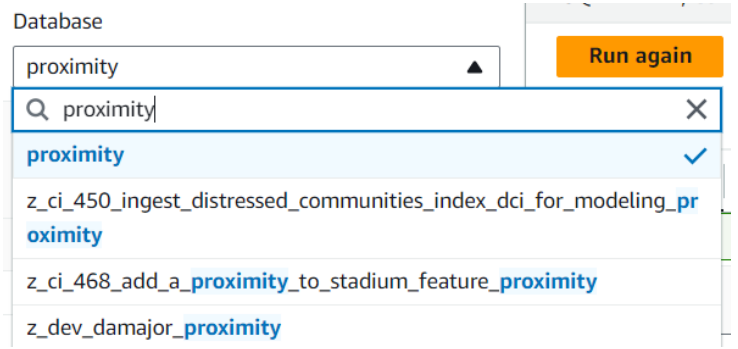
🤢 Table changes run manually

Better infrastructure

dbt

🚀 Continuous integration for data transformations

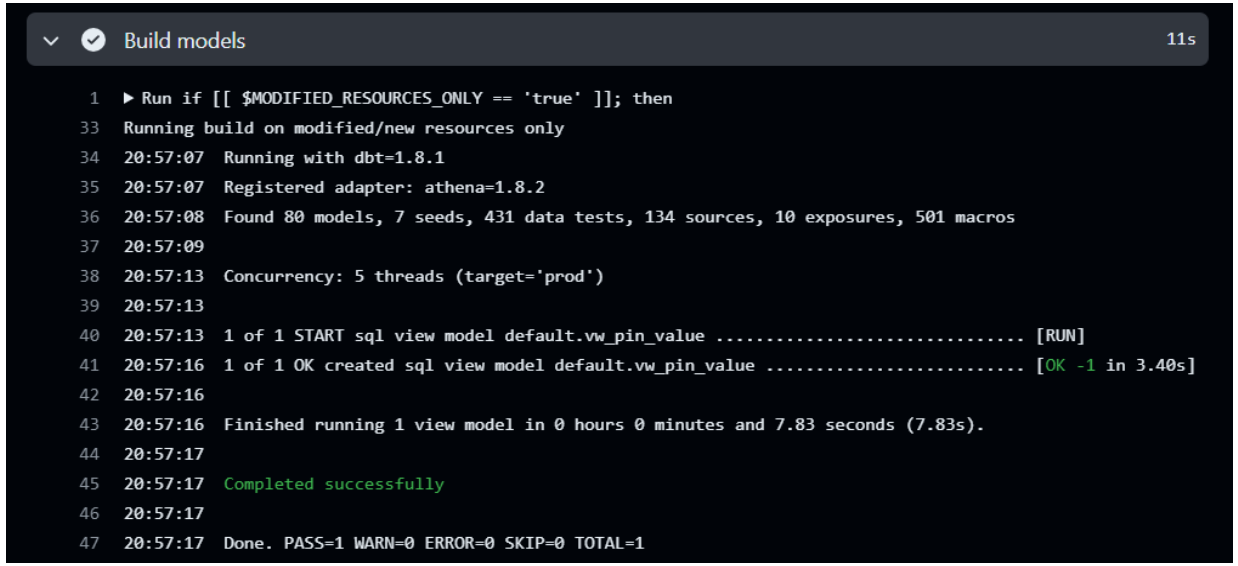
After...



💪 Dev/prod resources separated by schema prefixes

😎 New + modified resources Deployed by CI

👍 Prod resources protected From manual editing



Better infrastructure

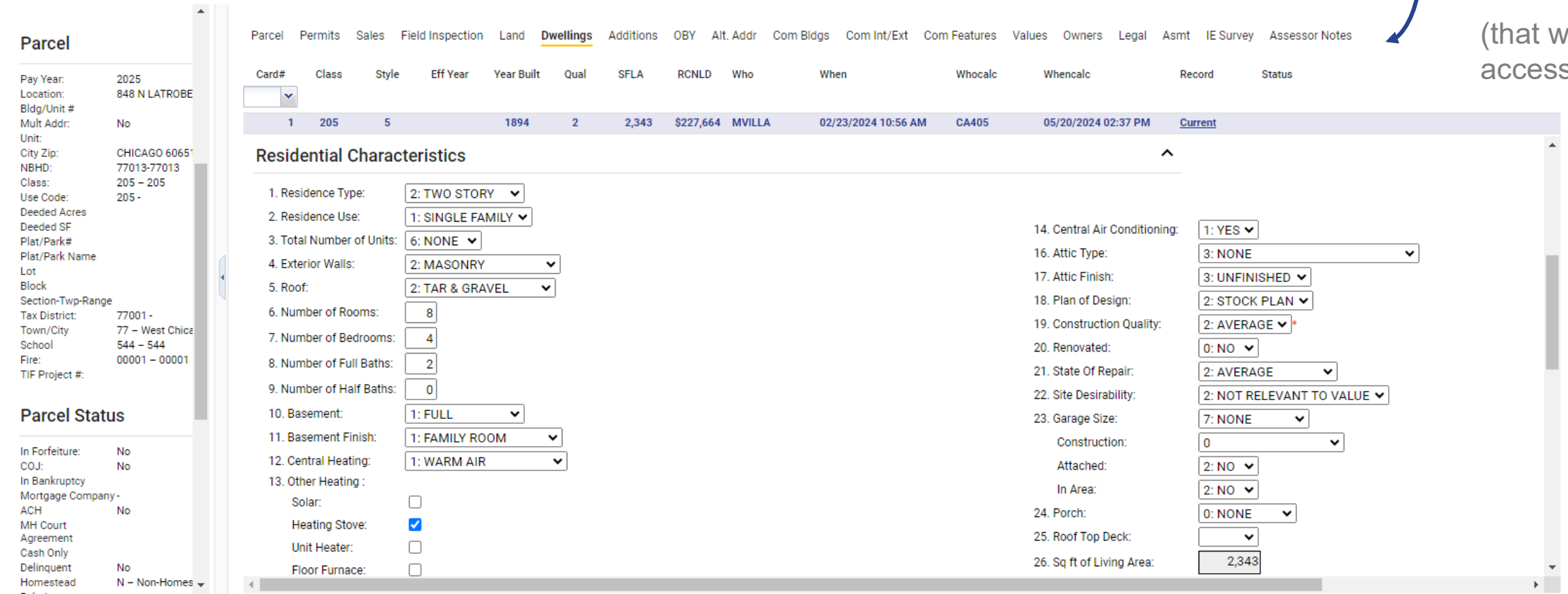
dbt

Data quality testing

- Data comes from vendor DB with little data validation

37 tables,
5,512 columns,
817m rows

(that we have access to)



Better infrastructure

dbt

🧪 Data quality testing

- Data comes from vendor DB with little data validation
- We define tests on dbt source definitions

433 tests

```yaml
      - name: sfla
        description: '{{ doc("shared_column_char_bldg_sf") }}'
        data_tests:
          - not_null:
              name: iasworld_dweldat_sfla_not_null
              additional_select_columns: *select-columns
              config: *unique-conditions
              meta:
                description: sfla (Building Square Footage) should not be null
          - accepted_range:
              name: iasworld_dweldat_sfla_between_1_and_2200_for_class_205
              min_value: 1
              max_value: 2200
              additional_select_columns: *select-columns-with-class
              config:
                where: |
                  CAST(taxyr AS int) BETWEEN
                    {{ var('test_qc_year_start') }} AND {{ var('test_qc_year_end') }}
                  AND cur = 'Y'
                  AND deactivat IS NULL
                  AND class = '205'
              meta:
                description: >
                  sfla (Building Square Footage) should be between
                  1 and 2200 for class 205 cards
```

Better infrastructure

dbt

🧪 Data quality testing

- Data comes from vendor DB with little data validation
- We define tests on dbt source definitions
- Python script parses test results for output to Excel

	A	B	E	F	I	AL
1	**Test description fields**		**Unique identifier fields**			
2	*These fields identify a failing test.*		*These fields identify the row that failed*			
3	source_table	description	taxyr	parid	class	sfla
10903	dweldat	sfla should be between 1 and 2200 for class 205 cards	2024	16043300200000	205	2343
10918	dweldat	sfla should be between 1 and 2200 for class 205 cards	2024	16013160100000	205	2670
10927	dweldat	sfla should be between 1 and 2200 for class 205 cards	2024	16041000220000	205	2881
10950	dweldat	sfla should be between 1 and 2200 for class 205 cards	2024	16041100390000	205	2569
10954	dweldat	sfla should be between 1 and 2200 for class 205 cards	2024	16043000160000	205	2357

data-architecture / .github / scripts / **transform_dbt_test_results.py**

Code | Blame | 1241 lines (1103 loc) · 48.4 KB

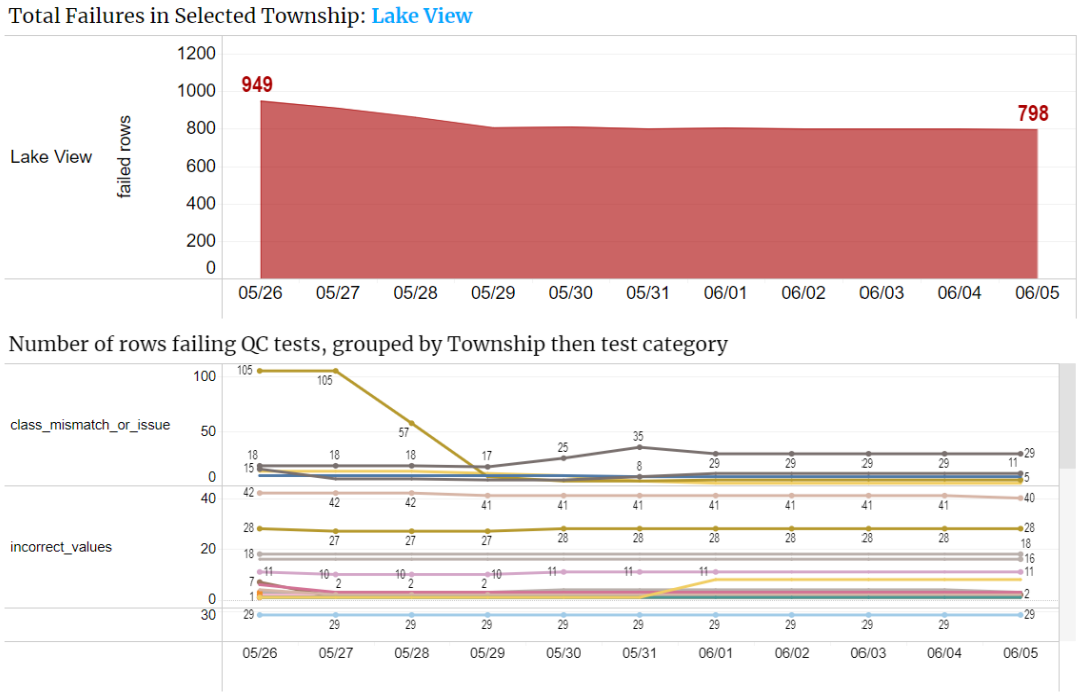
```
678
679  def main() -> None:
680      """Entrypoint to this script. Parses dbt test results and writes artifacts
681      to the output directory with metadata about tests."""
682      try:
683          run_results_filepath = sys.argv[1]
684      except IndexError:
685          run_results_filepath = os.path.join("target", "run_results.json")
686
687      try:
688          manifest_filepath = sys.argv[2]
689      except IndexError:
690          manifest_filepath = os.path.join("target", "manifest.json")
691
692      date_today = datetime.datetime.today().strftime("%Y-%m-%d")
693      try:
694          output_directory = sys.argv[3]
695      except IndexError:
696          output_directory = f"qc_test_results_{date_today}"
697
698      test_cache_path = get_test_cache_path(
699          run_results_filepath,
700          manifest_filepath,
701      )
702
```

Better infrastructure

dbt

🧪 Data quality testing

- Data comes from vendor DB with little data validation
- We define tests on dbt source definitions
- Python script parses test results for output to Excel
- Tableau dashboard displays test failures over time



Better infrastructure

dbt

🧪 Data quality testing

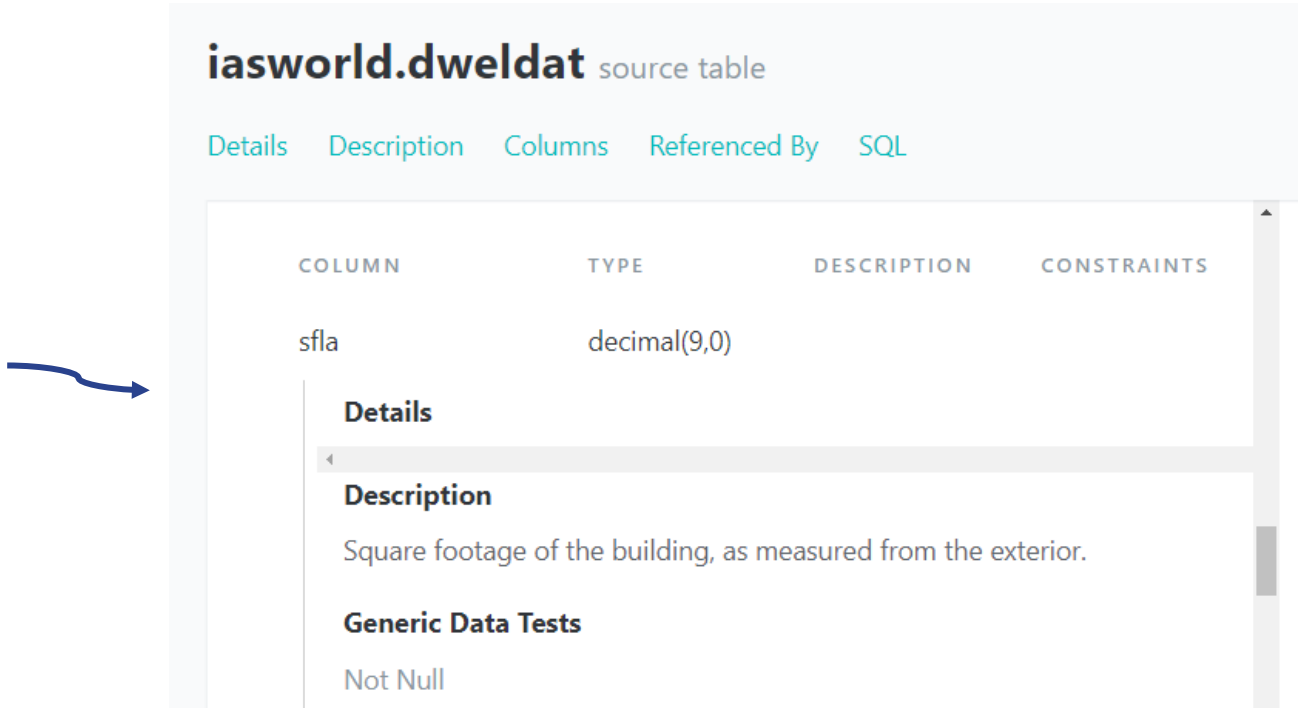
- Data comes from vendor DB with little data validation
- We define tests on dbt source definitions
- Python script parses test results for output to Excel
- Tableau dashboard displays test failures over time
- Eventually: Automate tests, catch errors in real time

Better infrastructure

dbt

📖 Data documentation

```
447
448   ## char_bldg_sf
449
450   {% docs shared_column_char_bldg_sf %}
451   Square footage of the building, as measured from the exterior.
452   {% enddocs %}
453
454   ## char_bldg_is_mixed_use
455
456   {% docs shared_column_char_bldg_is_mixed_use %}
457   The 10-digit PIN (building) contains a 14-digit PIN that is
458   neither class 299 nor 399.
459
460   Applies to condos only
461   {% enddocs %}
```



Taxpayer benefits

dbt

🚀 Continuous integration 👉 Faster and safer iteration

39 PRs merged last month

🧪 QC testing 👉 Fewer data errors that cause incorrect values

194 rows fixed last month

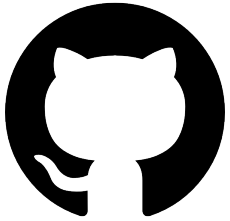
📖 Data documentation 👉 More durable knowledge, fewer mistakes

3,733 columns, tables, and tests documented

Better infrastructure

AWS Athena Serverless columnar data storage and retrieval

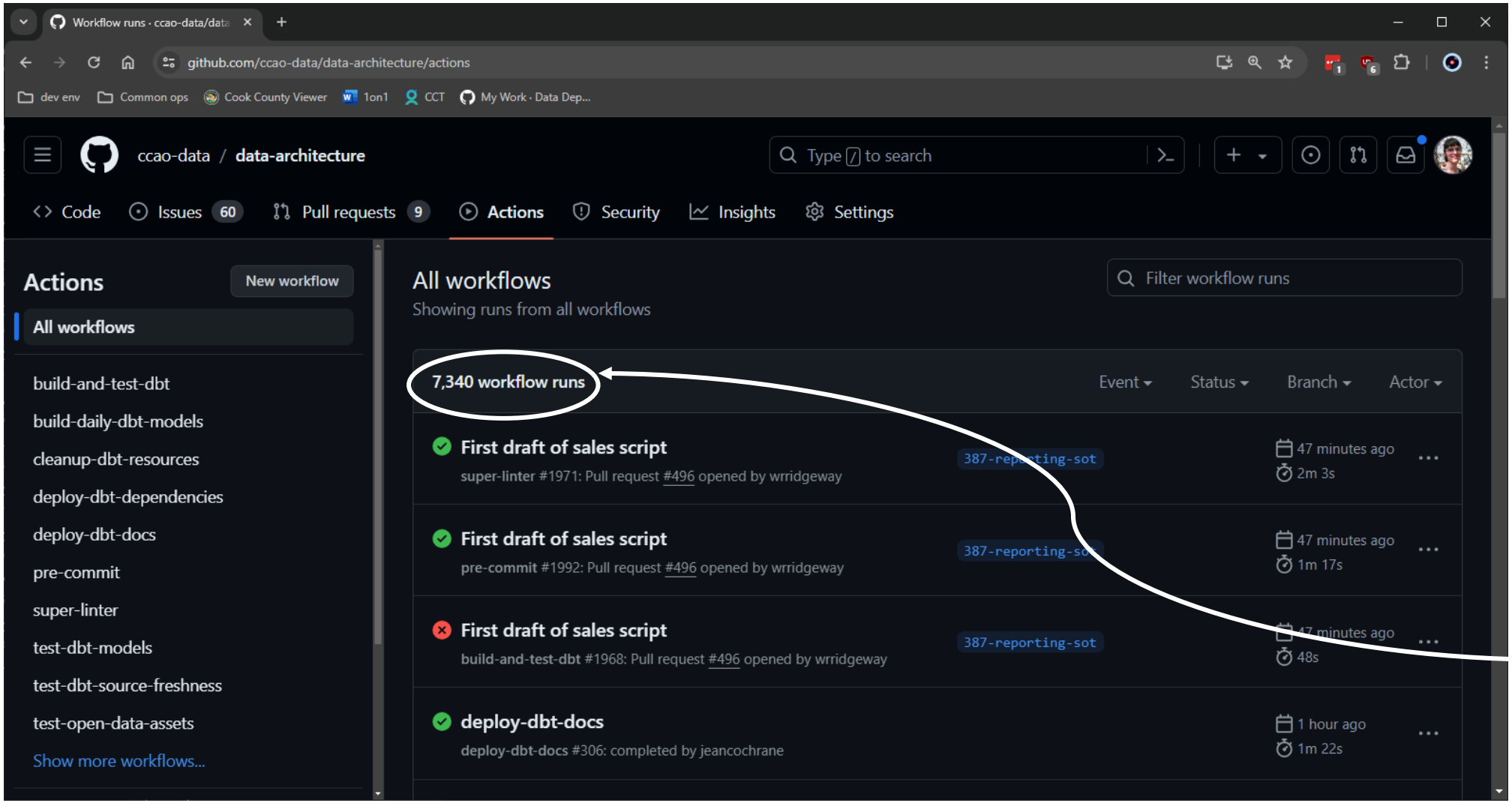
dbt Data transformation, testing, and documentation



GitHub Actions Code automation

Better infrastructure

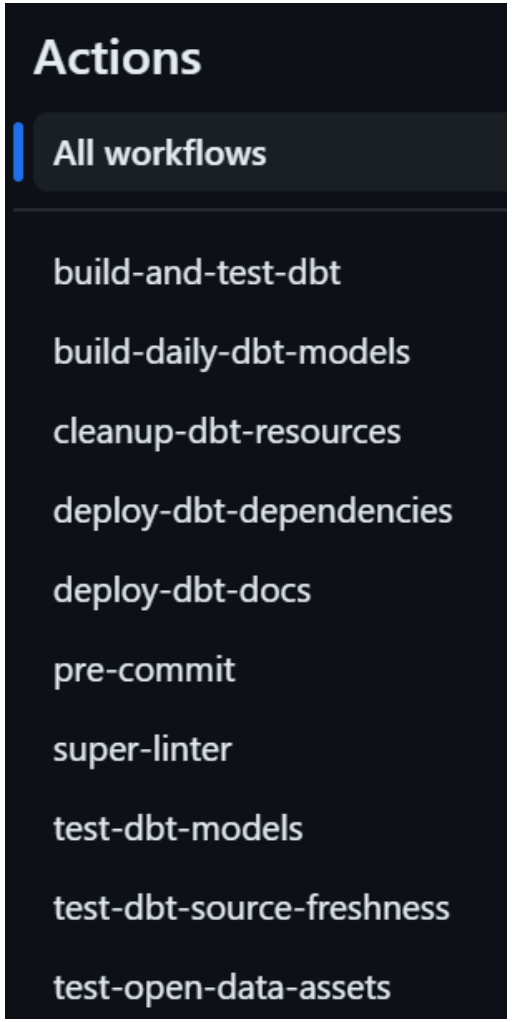
GitHub Actions



One year's worth of workflow runs

Better infrastructure

GitHub Actions



On commits

Build and unit test tables and views
Deploy Python dependencies for Spark
Deploy dbt docs

On PR close

Clean up staging resources

Scheduled daily

Refresh tables
Run QC tests
Run source freshness tests
Run open data asset tests

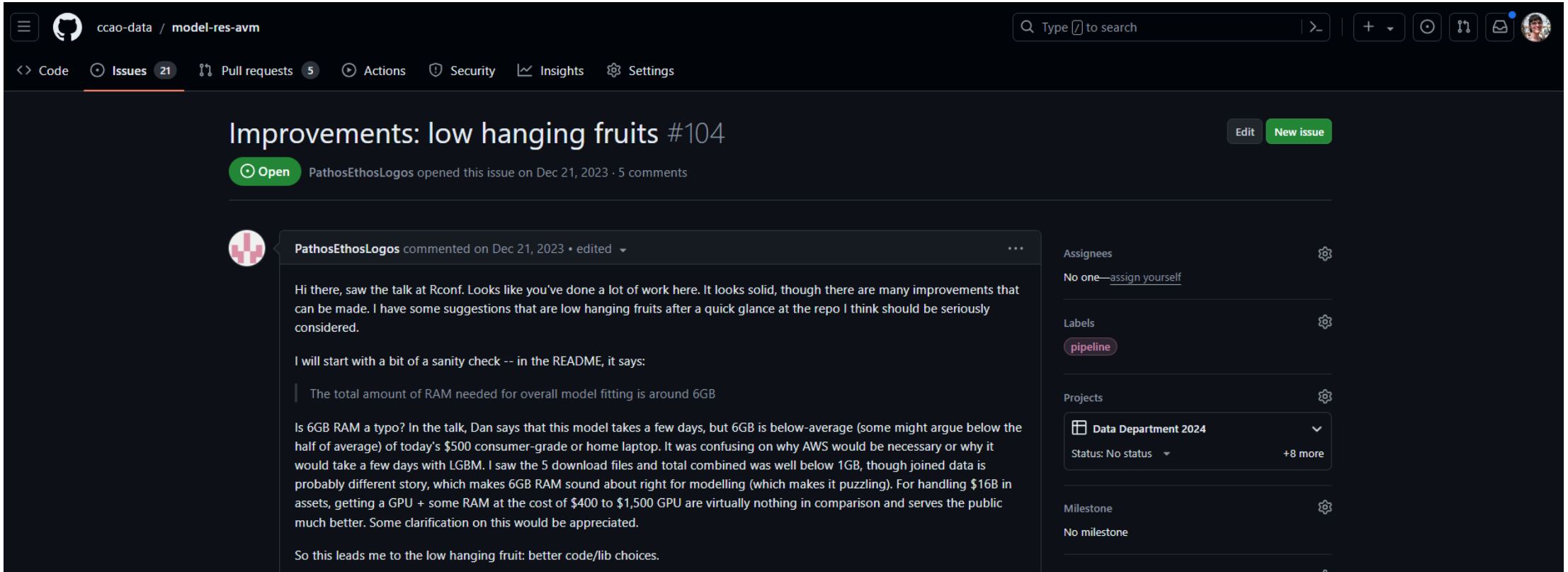
On dispatch

Build and run model
Delete model runs
Tag model runs

Taxpayer benefits

GitHub Actions

- ⏰ Saves **time**: Work on hard modeling problems, automate the rest
- 💸 Saves **money**: All features we use are on the free tier
- 📢 Builds **trust**: Most of our work is public and consolidated on one platform



Better infrastructure

Putting it all together: What's the impact of bad data?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		2024-02-06-relaxed-tristan												
3							Top 10 Most Important Model Variables							
4	meta_pin	meta_c	meta_class	pred_card_initial_fmv			meta_t	meta_r	char_bldg_	char_fb	char_yrblt	char_land_	char_frpl	loc_school_elementary_di
5	PIN	Card	Class	Original Prediction	API Prediction	Rounded API Prediction	Town	Nbhd	Bldg. Sqft.	F. Baths	Year Built	Land Sqft.	Fireplaces	Elementary District GEOID
20185	16-04-330-020-0000	1	205	$276,152			77	77013	843	2	1894	3528	0	609975
20186	16-04-330-021-0000	1	205	$198,675			77	77013	1752	2	1908	3150	0	609975
20187	16-04-330-022-0000	1	205	$196,718			77	77013	1752	2	1908	3150	0	609975
20188	16-04-330-023-0000	1	211	$200,977			77	77013	1763	2	1893	4725	0	609975
20189	16-04-330-024-0000	1	206	$323,456			77	77013	2348	2	1888	4725	0	609975
20190	16-04-330-025-0000	1	205	$222,424			77	77013	2156	1	1898	4725	0	609975
20191	16-04-330-026-0000	1	205	$185,047			77	77013	1828	2	1893	1575	0	609975
20192	16-04-330-027-0000	1	205	$203,473			77	77013	1828	2	1893	3150	0	609975
20193	16-04-330-028-0000	1	205	$211,909			77	77013	2112	1	1893	4095	0	609975
20194	16-04-330-030-0000	1	211	$241,766			77	77013	2022	2	1908	3150	0	609975
20195	16-04-330-031-0000	1	211	$351,889			77	77013	4126	4	1913	4725	0	609975
20196	16-04-330-032-0000	1	212	$375,279			77	77013	5280	6	1916	3601	0	609975
20197	16-04-330-035-0000	1	212	$201,372			77	77013	2600	4	1916	3200	0	609975
20198	16-04-330-040-0000	1	212	$206,213			77	77013	2822	3	1901	3121	0	609975
20199	16-04-331-001-0000	1	211	$341,460			77	77013	4248	4	1913	3150	0	609975
20200	16-04-331-003-0000	1	211	$247,267			77	77013	2394	2	1903	4725	0	609975
20201	16-04-331-004-0000	1	205	$185,155			77	77013	1248	1	1903	4725	0	609975
20202	16-04-331-005-0000	1	211	$226,464			77	77013	2208	2	1898	1575	0	609975
20203	16-04-331-006-0000	1	211	$247,208			77	77013	2208	2	1898	3150	0	609975

Thank you! Questions?

COOK COUNTY
ASSESSOR'S
OFFICE

Links

`@ccao-data/data-architecture`	ELT and data lake repo
`@ccao-data/model-res-avm`	Residential model repo
`@ccao-data/model-condo-avm`	Condo model repo
📖 **ccao-data.github.io/data-architecture**	Data documentation
💌 **assessor.data@cookcountyil.gov**	Team email address