



The Ronald and Maxine Linde Center  
for Science, Society, and Policy

Caltech

RESEARCH PROJECT BRIEFING

# Measures and Models for Voter List Accuracy in Orange County (CA)

---

*Progress Report*

PRESENTED BY

**R. Michael Alvarez · Alexander Halsey**

*California Institute of Technology*

September 2026

# Thank you!

---



## ORANGE COUNTY REGISTRAR OF **VOTERS**

At the outset, we want to thank our project partner, the Orange County Registrar of Voters Bob Page and Justin Berardino, OCROV Deputy Director, Operations.

- Thanks to MEDSL for supporting this project.
- Thanks to Jian Cao for his assistance.
- Thanks to the Caltech LCSSP for support.

# Project overview

# What the Project Does

---

- **Build improved measures and models that provide actionable voter file intelligence**

For large-county, PII-free voter files.

- **Use daily snapshots across election cycles**

To analyze voter file maintenance across two families of measures: file accuracy and threat indicators.

- **Develop ML and AI tools**

For improving voter list accuracy and maintenance, with easy-to-use dashboards for election officials.

- **Build comparative metrics across counties**

Using snapshots of the CA state voter file.

- **Continue to build our collaboration with OCROV and other election officials**

# Development: May to September 2026

---

01

MAY – JUNE

## Audit & Redesign

Audited past voter file analysis pipelines; built new approach using an expanded set of measures. Confirmed via synthetic voter file data, then real voter file data.

02

JUNE – AUGUST

## Refine in Consultation

Continued development and refinement of methodology and pipeline, in consultation with OCROV.

03

AUGUST – SEPTEMBER

## Production & Expansion

Production pipeline established for OC data; dashboard released for registrar use. Began porting to CA data and modeling.

# The Data

- **105 snapshots**

13 April to 11 September 2026; 104 consecutive pairs.

- **Tab-delimited countywide extracts**

Roughly 973 MB each, 62 recognized fields.

- **Intervals of one to four days**

Gaps are weekends, holidays, or one day not collected.

- **Active registrants only**

The county supplies no inactive records.

- **Secured-server processing**

Software development aided by Claude (usually Opus 5).

~1.9M

registrants per snapshot

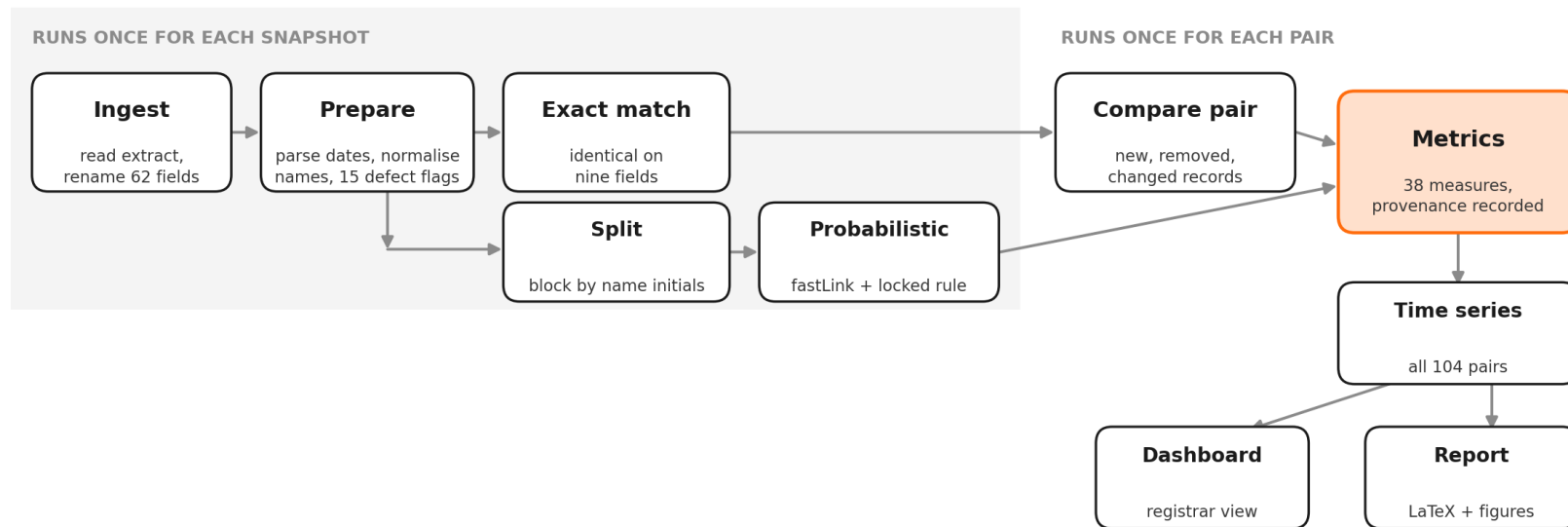
105

snapshots

62

fields tracked

# Pipeline Architecture



# Three Built-In Validation Checks

## 1 Within a run — every middle file is computed twice

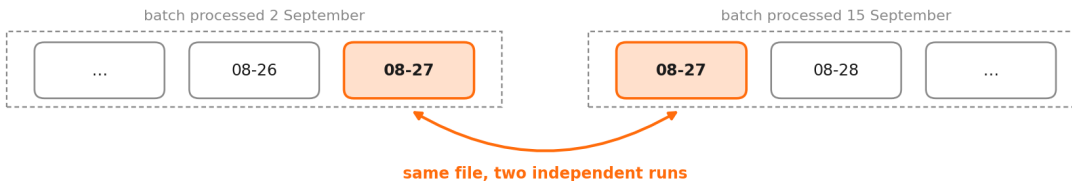
The audit always compares two consecutive files, so file 2 is the second file of pair 1 and the first file of pair 2.



0 disagreements of 103

## 2 Across deliveries — the bridge file, weeks apart

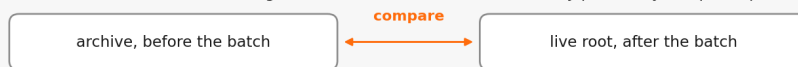
When the registrar sends the next batch, the last file of the old batch is processed again as the first file of the new one.



metrics agree exactly

## 3 Non-disturbance — adding data must not change old results

Before each batch runs the existing results are archived; afterwards every previously computed pair is compared against that archive.



95 of 95 pairs identical

# Measures & metrics

# Metrics I — File Accuracy

---

## ■ COMPLETENESS

Missing last name, first name, date of birth, house number, or residential ZIP.

## ■ MALFORMED

15 flags: whitespace in names, unparseable dates, invalid ZIP, unexpected codes.

Thirteen form the composite; a blank first name is allowed.

## ■ FORMAT VALIDITY

Birth date in the future, age over 110, registration date in the future.

All date-dependent measures anchored to the extract date, not the run clock.

## ■ LOGICAL CONSISTENCY & NAME QUALITY

Internal contradictions, placeholder values.

# Metrics II — Duplicates

---

## ■ Two independent detectors

Outputs overlap and must not be added.

## ■ EXACT

Agreement on all nine of: three name fields, three date-of-birth components, house number, street, ZIP.

## ■ PROBABILISTIC

fastLink linkage on six fields, then a locked post-filter requiring:

Full date of birth present and equal on both records.

Middle initial and suffix corroborate where present; blank is neutral.

Both surname and forename clearly different before a pair is dropped.

## ■ Counts are lower bounds by design

Blocking and the full-DOB rule both restrict recall and help prevent false positives.

# Metrics III — Change and Threat

---

## ■ CHANGE

Eleven fields compared between snapshots for records matched by identifier: names, date of birth, address, ZIP, party, status, precinct, status reason, absentee status.

A change is: exactly one side missing, or both present and different.

## ■ THREAT — High-Occupancy Addresses

More than 10 registrants at one normalized address.

## ■ THREAT — Record Volatility

Share of the baseline roll with any field change; broken down by age band, party group, and congressional district.

# Anomaly Detection

- **38 metrics monitored**

21 per snapshot, 17 per pair.

- **Robust band**

Median  $\pm 2 \times 1.4826 \times$  median absolute deviation.

- **Rolling window**

14-snapshot window for drifting measures; flat band for stable ones.

- **Multiplicity controlled**

Benjamini-Hochberg at  $\alpha = 0.05$  across all tests. Flagged events are subject to OCROV review.

**3,973**

statistical tests run

**327**

outside their band

**141**

significant after  
correction

# Recent results

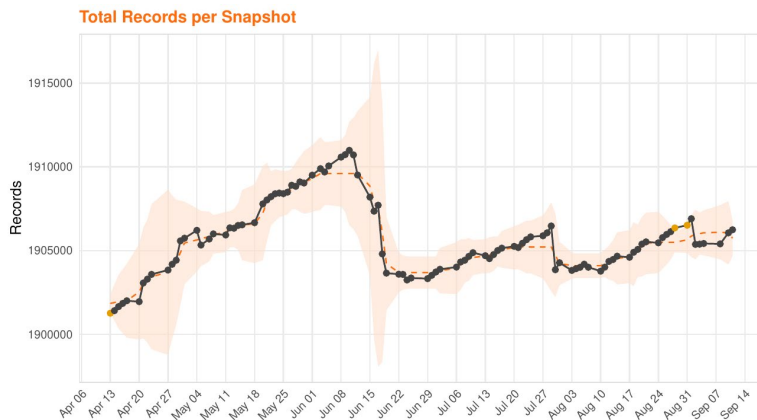
*Every panel below is generated directly by the reporting pipeline — these graphics form the basis for reporting shared with OCROV.*

## RECENT RESULTS

# The Roll and Its Defect Rate

*A stable dynamic and a stable defect rate — the baseline everything else is read against.*

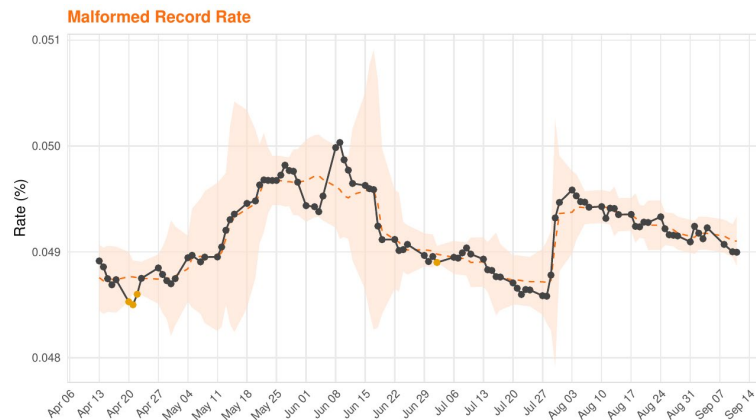
## TOTAL REGISTRANTS



~1.9M records throughout; mid-June shows a maintenance step.

*Shaded band is the usual range; dashed line the typical level; red points are flagged and survive multiplicity correction, amber flagged only.*

## MALFORMED RATE



0.049% malformed (~935 records/snapshot), no trend.

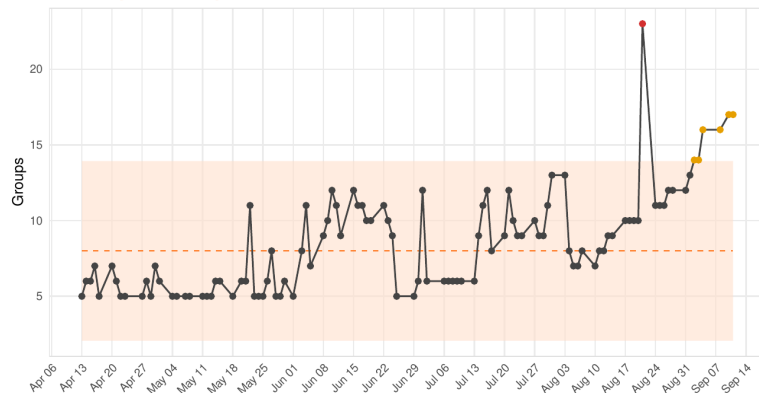
## RECENT RESULTS

# Duplicate Detection

*Duplicates are rare and relatively stable — we began seeing some change to the dynamic in recent data.*

## EXACT DUPLICATE GROUPS

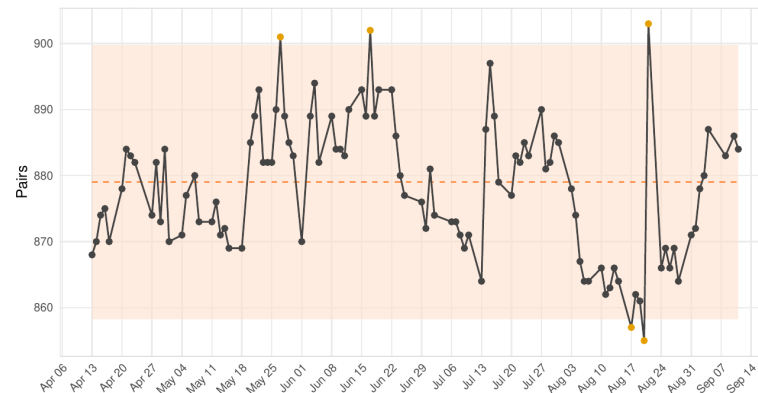
Exact Duplicate Groups



Median 8, range 5–17; every group observed was a pair.

## PROBABILISTIC DUPLICATE PAIRS

Near-Duplicate Pairs (Probabilistic)



Near-duplicate pairs, median 879 per snapshot.

*Combined duplicate rate 0.092% of records — the union of both detectors, which overlap and must not be added.*

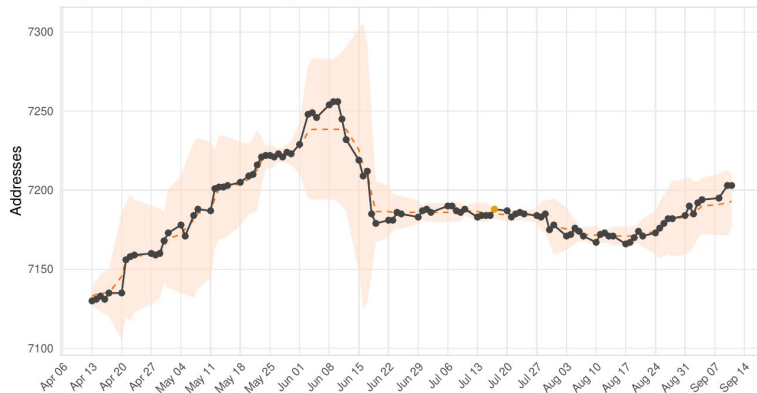
## RECENT RESULTS

# Threat Indicators

*High-occupancy addresses move in steps; record volatility stable with periodic spikes reflecting maintenance activities.*

## HIGH-OCCUPANCY ADDRESSES

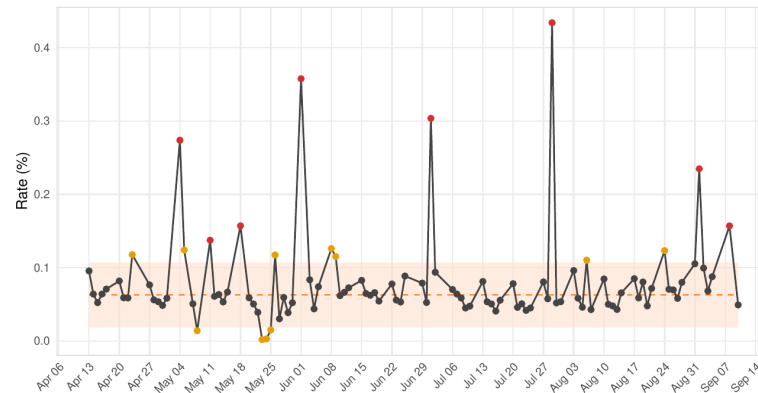
High–Occupancy Addresses (> 10 Registrants)



7,184–7,256 addresses; a clear step down in mid-June.

## RECORD VOLATILITY

Record Change Volatility (% of roll)



Median 0.063% of the roll; max 0.43% on 27 July.

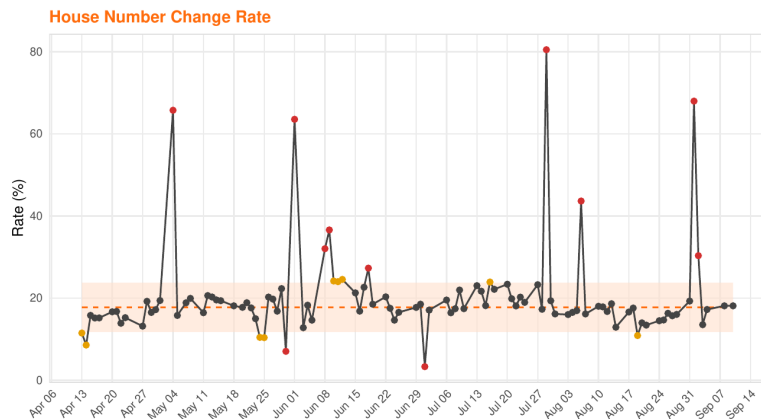
*The address key omits unit numbers by design, so apartment buildings are expected to appear — the level says little, the series tells the story.*

## RECENT RESULTS

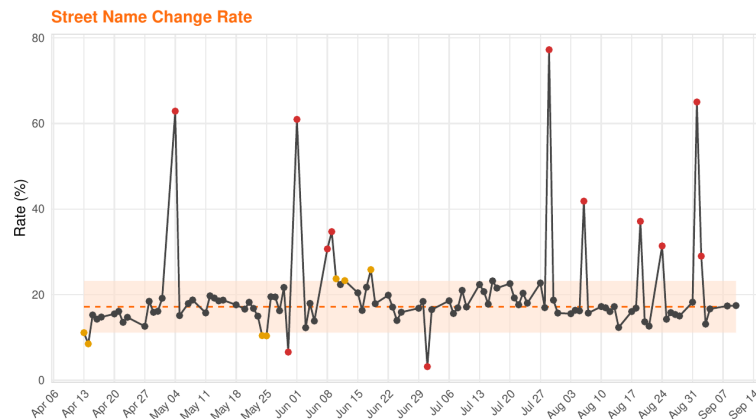
# One Event, Two Measures

*Relatively stable with some spikes reflecting maintenance activities.*

## HOUSE NUMBER CHANGE RATE



## STREET NAME CHANGE RATE



*Three independent measures agreeing on one date is the strongest evidence the series detects real events.*

# Party Registration Change

- **One in five edited records involves a party change**

Most involve No Party Preference.

- **Typically ~1,300 records change per file-pair**

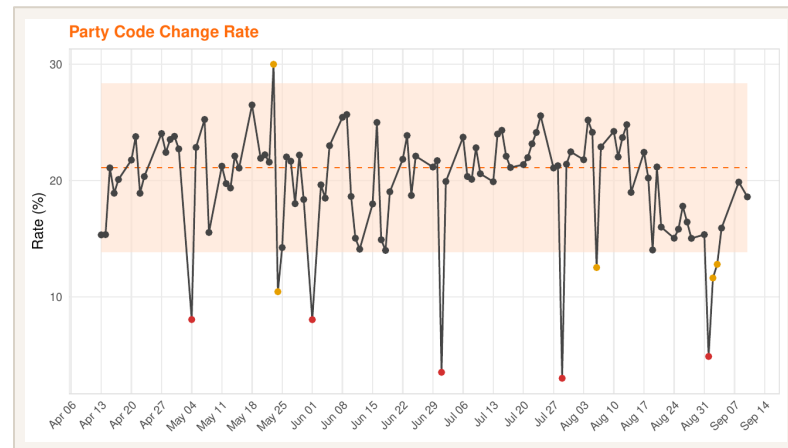
Ranging from 40 to 8,400 depending on office activity; ~21% of those edits involve a party change.

- **Across the full series: 1.4% of 1.95M registrants**

Changed party at least once in six months; 69% of that movement has No Party Preference on one side.

- **Major-party movement nearly cancels out**

1,686 Republicans became Democrats; 1,727 Democrats became Republicans — a net of 41 people.



*Party code change rate among edited records, over the full series.*

# Interactive dashboard

*Snapshots of the fully interactive dashboard provided to OCROV*

# INTERACTIVE DASHBOARD

## Voter File Auditing

### Monitor

- Tier 1 metrics
- Explore metrics
- Compare updates
- Anomalies

### Explore

- Demographic breakdowns**
- Methods

Signed in as [ahalsey@caltech.edu](#)

Sign out

## View options

### Dimension

Age band

### Metric

Registrants (n)

Request a review

## Demographic breakdowns

Compare aggregate measures across age bands, party groups, or congressional districts. Where applicable, measures can be shown as a percentage of total registrants in a given demographic group.

### Registrants (n) by age band

#### Trends by group

Shown as raw counts. Each panel uses its own scale. See the composition chart below for each group's share of the total.



## INTERACTIVE DASHBOARD

### Compare updates

Track how much changes between consecutive voter file snapshots, or inspect a single comparison in detail — fields updated, status changes, and party movement. Use the view options in the sidebar to choose a measure and date range, or to pick the comparison to inspect.

View

Across updates

Inspect one comparison

Latest comparison

0.0493% (940 records)

Change from prior comparison

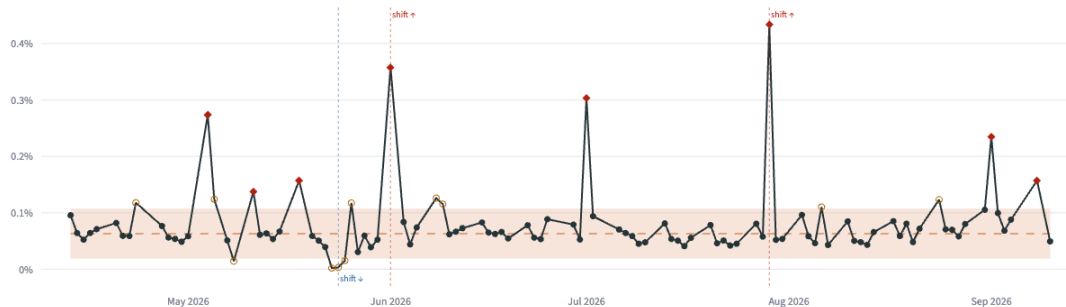
-0.108 pp (-2,052 records)

Current status 

Within recent range

### Record volatility

Share of the earlier snapshot's records that were changed by the later snapshot.



> Automated flag history for this measure (19)

INTERACTIVE DASHBOARD

Voter File Auditing

Monitor

- Tier 1 metrics
- Explore metrics
- Compare updates
- Anomalies

Explore

- Demographic breakdowns
- Methods

Signed in as [ghabreu@caltech.edu](#)

Sign out

Tier 1 metrics

Summary of highest priority statistics. Select the checkbox next to a row to see more detail.

Data through  
Sep 11, 2026

Registered voters  
1,906,249

Latest comparison  
Sep 11, 2026

Total registered voters in each snapshot.

Threat indicators

Address concentration

| Measure                                 | Latest | Since previous | Recent trend | Current status      |
|-----------------------------------------|--------|----------------|--------------|---------------------|
| Addresses with > 10 registrants         | 7,203  | +0             |              | Within recent range |
| Registrants at high-occupancy addresses | 294K   | +59            |              | Low priority        |

Changes between files

| Measure           | Latest                | Since previous            | Recent trend | Current status      |
|-------------------|-----------------------|---------------------------|--------------|---------------------|
| Record volatility | 0.0493% (940 records) | -0.108 pp (2,052 records) |              | Within recent range |

File accuracy

Missing information

| Measure                 | Latest        | Since previous         | Recent trend | Current status                      |
|-------------------------|---------------|------------------------|--------------|-------------------------------------|
| Last name missing       | 0 (0%)        | +0 records (+0 pp)     |              | Within recent range                 |
| DOB missing             | 0 (0%)        | +0 records (+0 pp)     |              | Within recent range                 |
| House number missing    | 624 (0.0327%) | +0 records (+0.001 pp) |              | >5 records below last month's level |
| Residential ZIP missing | 0 (0%)        | +0 records (+0 pp)     |              | Within recent range                 |

Duplicate records

| Measure                    | Latest          | Since previous         | Recent trend | Current status                       |
|----------------------------|-----------------|------------------------|--------------|--------------------------------------|
| Combined duplicate records | 1,756 (0.0921%) | -4 records (-0.001 pp) |              | Within recent range                  |
| Exact duplicate records    | 34              | +0                     |              | >24 records above last month's level |

Record formatting

| Measure           | Latest       | Since previous         | Recent trend | Current status      |
|-------------------|--------------|------------------------|--------------|---------------------|
| Malformed records | 934 (0.049%) | +0 records (+0.001 pp) |              | Within recent range |

Automated flag history (47 signals)

LCSSP

The Ronald and Maxine  
Linde Center for Science,  
Society, and Policy

Caltech

# INTERACTIVE DASHBOARD

## Voter File Auditing

Monitor

Tier 1 metrics

Explore metrics

Compare updates

Anomalies

Explore

Demographic breakdowns

Methods

Signed in as [ahalsey@caltech.edu](#)

Sign out

## View options

Category

Accuracy — Duplicate records

Measure

Near-duplicate pairs (probabilistic)

Start date

2026/04/13

End date

2026/09/11

☒ Show usual range

☒ Show election dates

## Explore metrics

Follow a measure across voter file snapshots. Each measure's usual range and flagged values (potential anomalies) are shown, as well as election dates. Use the view options in the sidebar to view different measures or toggle chart settings.

Latest  
884

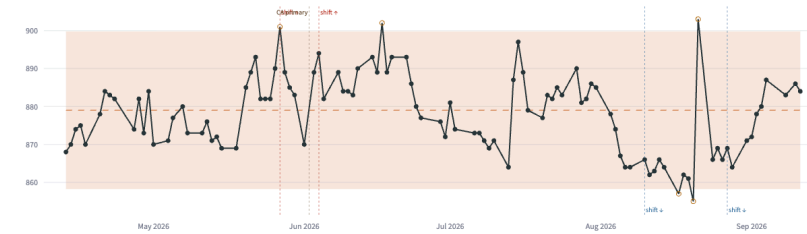
Since previous file  
-2

Current status  
Within recent range

### Near-duplicate pairs (probabilistic)

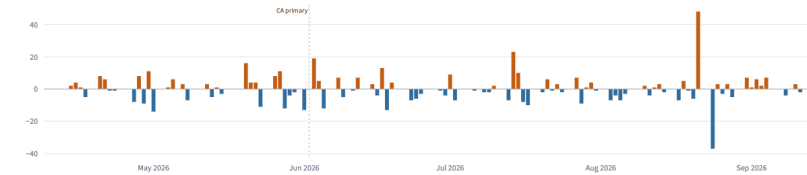
Pairs of records flagged as likely duplicates by the probabilistic matcher. A single record can appear in multiple pairs.

Level over time



### Change since the previous file

Change snapshot-over-snapshot.



# To-dos

*Short- and long-term deliverables*

# Next Steps

---

## CALIFORNIA STATEWIDE

Bring the state pipeline into line: seeding, anchoring, provenance, reporting.

- Filter the state file to active registrants so the two are comparable.
- Produce county rankings; validate against OC's own figures.

## NOVEMBER GENERAL

Completes a second cycle, allowing a party-gated primary to be compared against an ungated one.

## PARTY SWITCHING

Full analysis once two cycles are in hand.

## RESIDENTIAL ADDRESSES > 10 VOTERS

Detailed analyses underway.

## FASTER PROBABILISTIC MATCHING

Implement GPU-based methods.

## MODELING & DIMENSION REDUCTION

Person-period models of switching and record change; metrics as predictive factors for outcomes like ballot curing and undeliverable mail. Develop methods for reducing the metric dimensionality.

# Longer-term

---

- **Focus on fully-actionable metrics and measures**

Everything should be readily understandable, transparent, and actionable.

- **Run at any scale**

The same pipeline should work efficiently from a 50,000-voter county to a statewide file.

- **Open source the pipeline and dashboard**

Hand off to OCROV and other election officials for 2028.

- **Pre-registered cycle monitoring**

State in advance what a well-run file should look like through an election cycle, then publish what actually happened — turning auditing into a standard for election integrity.

- **Beyond California**

The methods are jurisdiction-agnostic. A multi-state consortium on a common metric set would allow for county and state comparisons.

# LCSSP

---

Linde Center for Science, Society, & Policy

*California Institute of Technology*

## CONNECT WITH US

**Web** <https://lcssp.caltech.edu/>

**Email** [lcssp@hss.caltech.edu](mailto:lcssp@hss.caltech.edu)

**Phone** 626-395-4228

**Social** [@CaltechLCSSP](#)



*Scan for resources*

# Thank you.

*Questions welcome.*