

The background of the slide is a photograph of a university campus. In the center, there is a bronze statue of a man in a long coat, standing on a stone pedestal. The statue is surrounded by lush green trees. In the background, there is a large, light-colored building with classical architectural features, including arched windows and columns. The overall scene is bright and sunny, with green foliage dominating the foreground and middle ground.

DATA TESTING WORKFLOW AUDIT

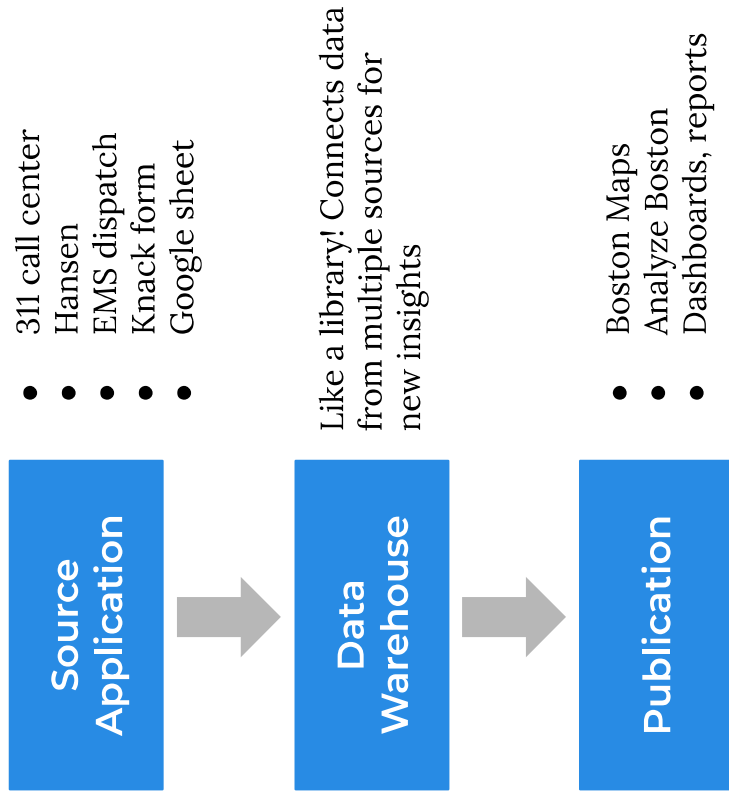
Jenna Flanagan

Project Scope

I performed an inventory and audit of 83 public-facing data workflows to improve data quality and alert on processing errors



WHAT IS A “DATA WORKFLOW”?



WHAT IS A “DATA UNIT TEST”?

The idea is borrowed from the software engineering version of unit testing, in which you try to make sure that all the individual parts, or “units” of your code work as they should before moving on to more complex testing of how the parts work together.

Similarly, unit testing of data involves asking questions about the quality of the data before you try to use it in a project.

You might think of Data Unit Testing like an alarm system for warning that data problems have cropped up.

Generally, Data Unit Testing is the process of asking some number of true/false questions about data. We can find out things like:


- Was this data updated within the last X days(weeks, months)?
- Does this required field contain any null values?
- Do latitude/longitude values fall in a range that makes sense for Boston locations?
- Does this field contain only the 5 categories we expect to see?
- Does this table contain the same number of columns it had previously?

WHEN AUTOMATION MESSES UP

DATA WORKFLOWS

EASY ERROR	MEDIUM ERROR	DIFFICULT ERROR
EASY TO SEE <ul style="list-style-type: none">• Access Failure/source down for maintenance• Warehouse error/destination down for maintenance• Alerting already built in to these processes• Human retries when system is available	MISPLACED DATA <ul style="list-style-type: none">• Naive computers may not see this as a failure• Google especially returns “empty file” rather than “not found error”• Computer will import a whole lot of nothing and report “success”!• DATA TESTING<ul style="list-style-type: none">• Look for empty table after import and send alert	“BAD” DATA <ul style="list-style-type: none">• Typos, sneaky spaces, and weird numbers, oh my• Can cause duplicate categories in a dashboard• Map location errors (0,0 is an actual point but has nothing to do with Boston)• ALERT FATIGUE<ul style="list-style-type: none">• Data testing can find these errors, but fixing may require stakeholder input

SAMPLE TESTING REPORT



great_expectations

Home

/ lagan_311_internal_data-sidewalk_repair_requests_vw / 429102985 / 2022-07-

25T17:04:36.171609+00:00 / d5b06af9c1a46f0e3df36f56c033df51

Expectation
Validation
Result

Evaluates whether a batch of data matches expectations.

Actions

Validation Filter:

Overview

Expectation Suite: [lagan_311_internal_data-sidewalk_repair_requests_vw](#)

Status: ✖ Failed

Statistics

Evaluated Expectations	13
Successful Expectations	11
Unsuccessful Expectations	2
Success Percent	≈84.62%

[Show more info](#)



SAMPLE TESTING REPORT, SUCCESS



Table-Level Expectations

Search

Status	Expectation	Observed Value
	Must have more than 1 rows.	3478

SAMPLE TESTING REPORT, SUCCESS

case_enquiry_id

Status	Expectation	Observed Value
✓	values must never be null.	100% not null
✓	values must be unique.	0% unexpected

case_status

Status	Expectation	Observed Value
✓	values must never be null.	100% not null
✓	distinct values must belong to this set: Open Closed .	['Closed', 'Open']

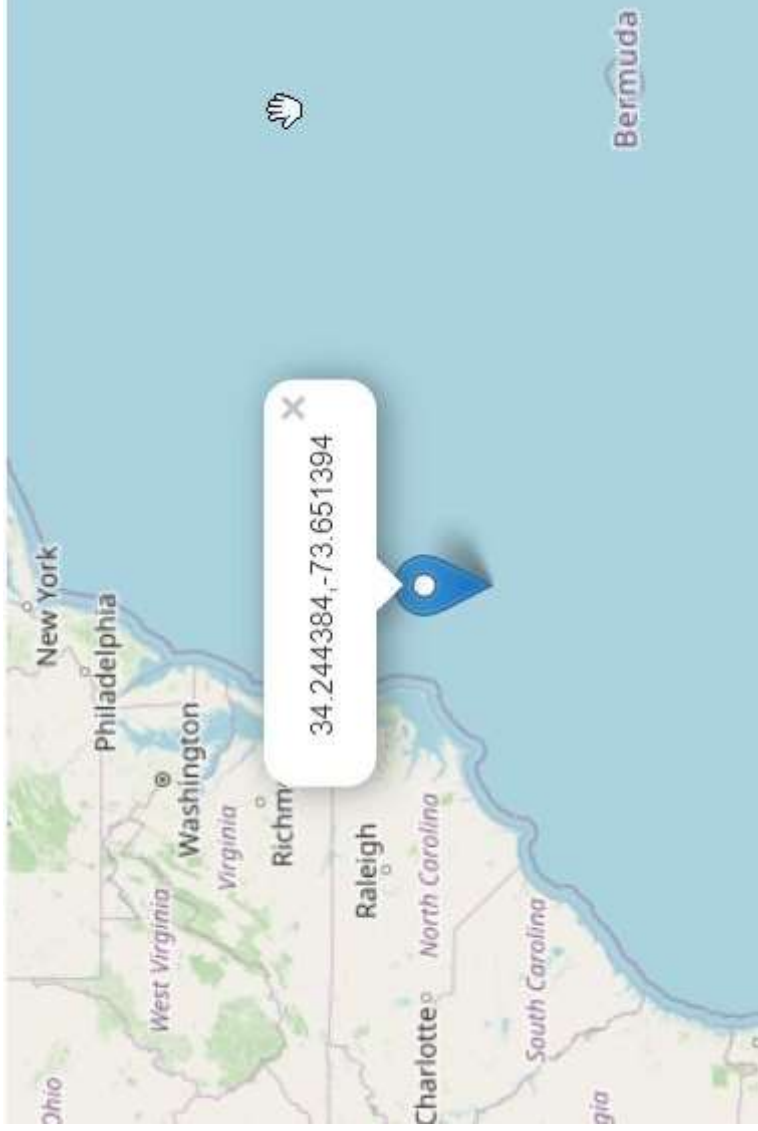
SAMPLE TESTING REPORT, ... WHAT'S THAT?

X

Search

Status	Expectation	Observed Value										
✓	values must never be null.	100% not null										
✗	values must always be between -71.2 and -71.0 .	≈0.92007% unexpected										
	32 unexpected values found. ≈0.9201% of 3478 total rows.											
	<table><thead><tr><th>Unexpected Value</th><th>Count</th></tr></thead><tbody><tr><td>-73.65139409812826</td><td>17</td></tr><tr><td>-73.65139090347309</td><td>9</td></tr><tr><td>-70.99916000133429</td><td>2</td></tr><tr><td>-70.99971000150636</td><td>1</td></tr></tbody></table>		Unexpected Value	Count	-73.65139409812826	17	-73.65139090347309	9	-70.99916000133429	2	-70.99971000150636	1
	Unexpected Value		Count									
	-73.65139409812826		17									
-73.65139090347309	9											
-70.99916000133429	2											
-70.99971000150636	1											

DATA... OOPSIE



The short story:

- Data underlying this view arrived with 0,0 for location information.
- Possibly people were using 0's to indicate "no information available"
- SQL query "transformed" this data into a map point that makes no sense.
- Computer understands 0 as a valid value. There is a special value for 'nothing' called "NULL".
- If we want the map to display correctly, using "NULL" is better than 0!

NOW WHAT?

DATA TEAM WORK

- Locate stakeholders
- Find out how data is being used (map? Dashboard? Just posted to Analyze Boston?)
- Can stakeholders fix easily?
- Can data team create an adjusted copy?

VISION/VALUES QUESTIONS

- How much work is this to fix?
- How much value will we add?
- Is this problem really worth having a daily “failure” report?
- How else do we not forget it’s on our to-do list?

OUTCOMES

DATA TESTING IMPROVEMENTS

OVERALL

- 83 workflows reviewed
- 33 already had tests
- 12 revised or archived by others
- 38 needed tests
- 76 tests created (some workflows needed multiple tests)
- Biggest workflow: 21 tests needed

DOCUMENTS

- Inventory spreadsheet with links
- Discovered notes feature to comment test code
- Learning about test suite documentation features

CLEANUP

- Editing geometry queries
- Tidying data categories
- 165 existing tests updated to add non-empty table test
- Vision conversation started!

Questions?

