

CRUSHING PIPELINE DEBT WITH GREAT EXPECTATIONS

Taylor Miller @ Superconductive
<http://tiny.cc/udem-ge>



I'm from Superconductive.com

We are data mercenaries for hire hell-bent on data quality.

AGENDA

1. A thing we do that is absolutely crazy
2. How to crush pipeline debt
3. Live demo



Here's some software!

BTW it's...



Here's some software!

BTW it's...

UNDOCUMENTED



Here's some software!

BTW it's...

**UNDOCUMENTED
UNTESTED**



Here's some software!
BTW it's...

**UNDOCUMENTED
UNTESTED
UNSTABLE**

A grayscale portrait of a person's face, which has been heavily textured with a wood-grain pattern, giving it a natural, organic feel. The person has short hair and is looking slightly to the right.

Here's some ~~software~~ data!

BTW it's...



Here's some ~~software~~ data!

BTW it's...

UNDOCUMENTED



Here's some ~~software~~ data!

BTW it's...

**UNDOCUMENTED
UNTESTED**



Here's some ~~software~~ data!

BTW it's...

**UNDOCUMENTED
UNTESTED
UNSTABLE**

PIPELINE DEBT



PIPELINES ARE LIKE SOFTWARE. AND ALSO NOT.

	SOFTWARE	PIPELINES
inputs are	usually known	often unknown
assumptions are	often crisp	murky
failures created by your code		data creators
tests verify	code	data
tests run	at compile time	at run time

BUT WHY TEST PIPELINES?

1. Risk mitigation
2. Pager mitigation
3. Increase trust & credibility
4. Codify knowledge & assumptions

PIPELINE RISKS:

MISSING / EMPTY FILES MANGLED LOADS

CORRUPTED DATA

SCHEMA CHANGES OUTAGES

UNEXPECTED VALUES

DATA RISKS:
DISTRIBUTION DRIFT
MODEL ASSUMPTIONS
EVOLUTION EDGE CASES
BIAS OUTLIERS

QUESTION TIME



GREAT EXPECTATIONS



ALWAYS KNOW WHAT TO EXPECT FROM YOUR DATA

EXPECTATIONS ARE ASSERTIONS ABOUT DATA

```
» expect_file_size_to_be_between  
» expect_table_row_count_to_be_between  
» expect_column_to_exist  
» expect_column_values_to_not_be_null  
» expect_column_values_to_be_unique  
» expect_column_values_to_be_between  
» expect_column_values_to_be_in_set  
» expect_column_values_to_match_regex  
» expect_column_mean_to_be_between  
» expect_column_kl_divergence_to_be_less_than  
» ... and many many more1
```

¹ https://docs.greatexpectations.io/en/latest/expectation_glossary.html

WHERE'S DOES THE COMPUTE HAPPEN?

GREAT EXPECTATIONS USES DIFFERENT BACK-END COMPUTE ENGINES

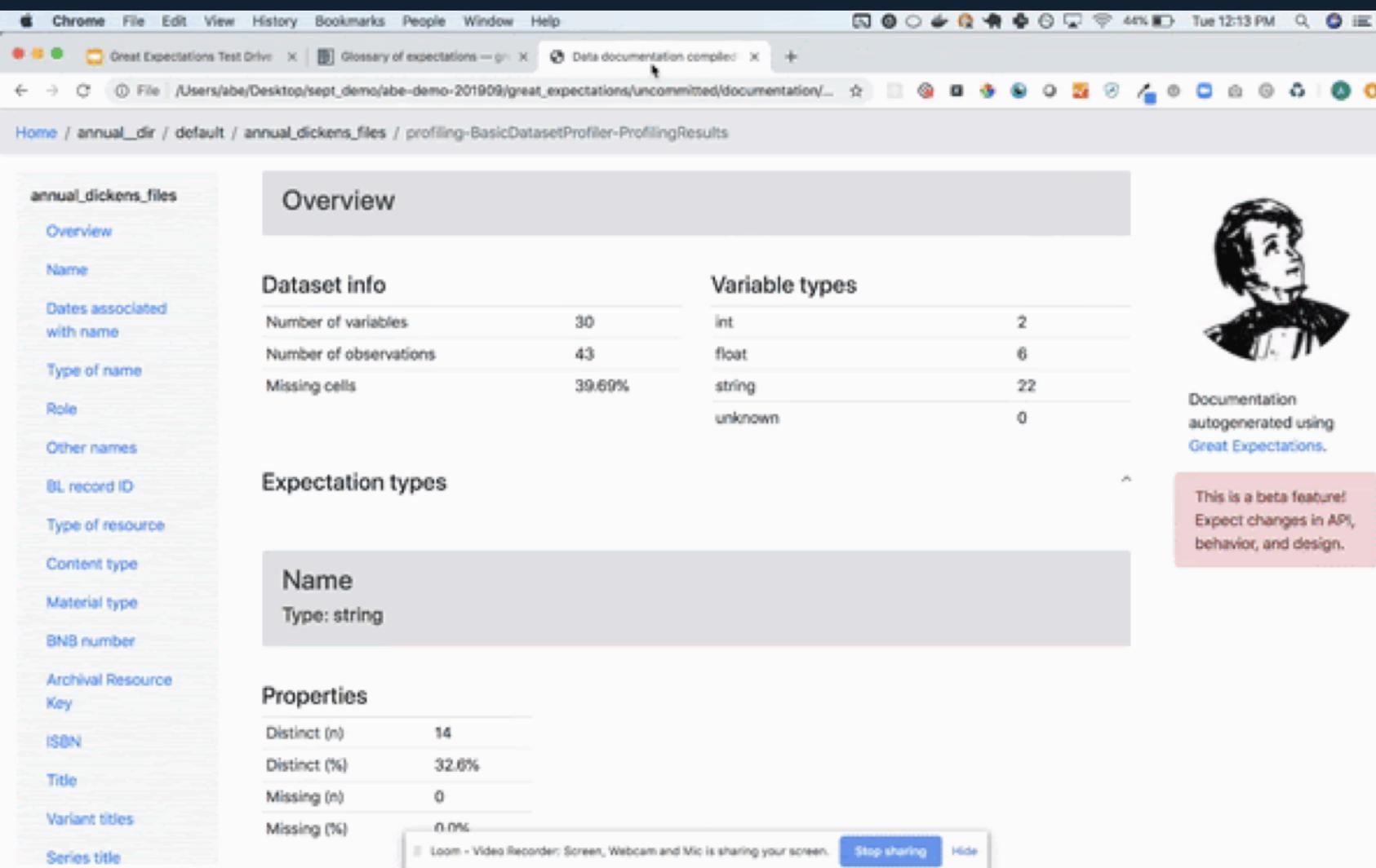
- » pandas
- » Spark
- » Common relational databases via SQLAlchemy
 - » Redshift
 - » BigQuery
 - » Snowflake
 - » Postgres
 - » MySQL

TESTS ARE DOCS

AND

DOCS ARE TESTS

TESTS ARE DOCS AND DOCS ARE TESTS



The screenshot shows a web browser window with several tabs open. The active tab displays a dataset overview for 'annual_dickens_files'. The page includes sections for 'Dataset info', 'Variable types', 'Expectation types', and 'Properties'. A sidebar on the left lists various metadata fields like Name, Dates associated with name, Type of name, Role, Other names, BL record ID, Type of resource, Content type, Material type, BNB number, Archival Resource Key, ISBN, Title, Variant titles, and Series title. A small portrait of a man is displayed next to the dataset information, along with a note about documentation being autogenerated using Great Expectations. A beta feature notice is also present. At the bottom, there's a message from Loom indicating screen sharing.

Dataset info		Variable types	
Number of variables	30	int	2
Number of observations	43	float	6
Missing cells	39.69%	string	22
		unknown	0

Name
Type: string

Properties	
Distinct (n)	14
Distinct (%)	32.6%
Missing (n)	0
Missing (%)	0.0%

- » Everything is JSON 
- » Compile to HTML or Notebooks 
- » Docs cannot get stale!

HOW TO CRUSH PIPELINE DEBT

DATA INGEST

especially if the data
isn't controlled by
your team

BEFORE & AFTER ML MODELS

prevent malfeasant AI

ANALYTIC WAREHOUSES

Be good data driven.

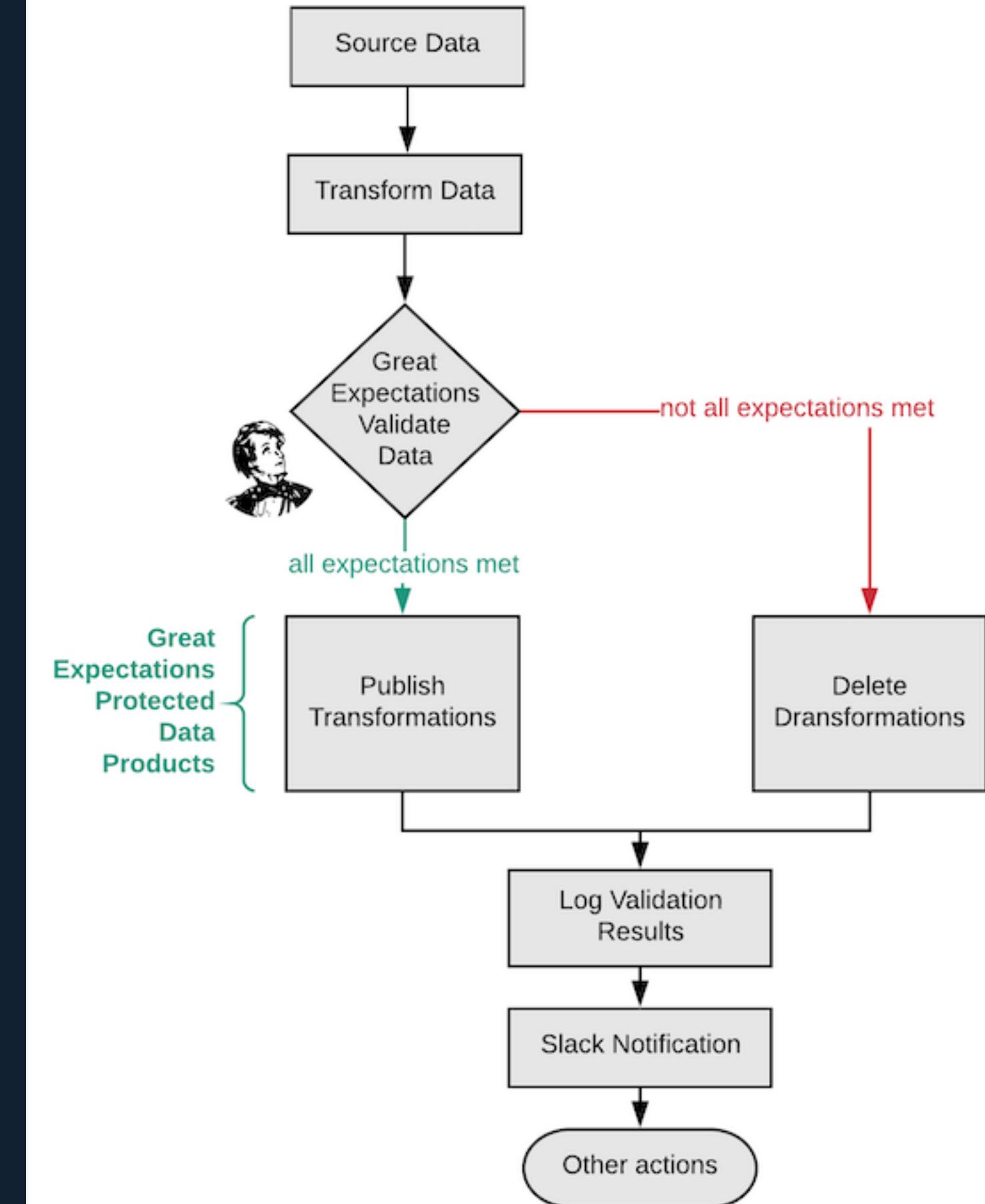
CRITICAL DASHBOARD TABLES

Don't piss off an
executive

HOW TO USE GREAT EXPECTATIONS IN A PIPELINE?

The WAP pattern!²

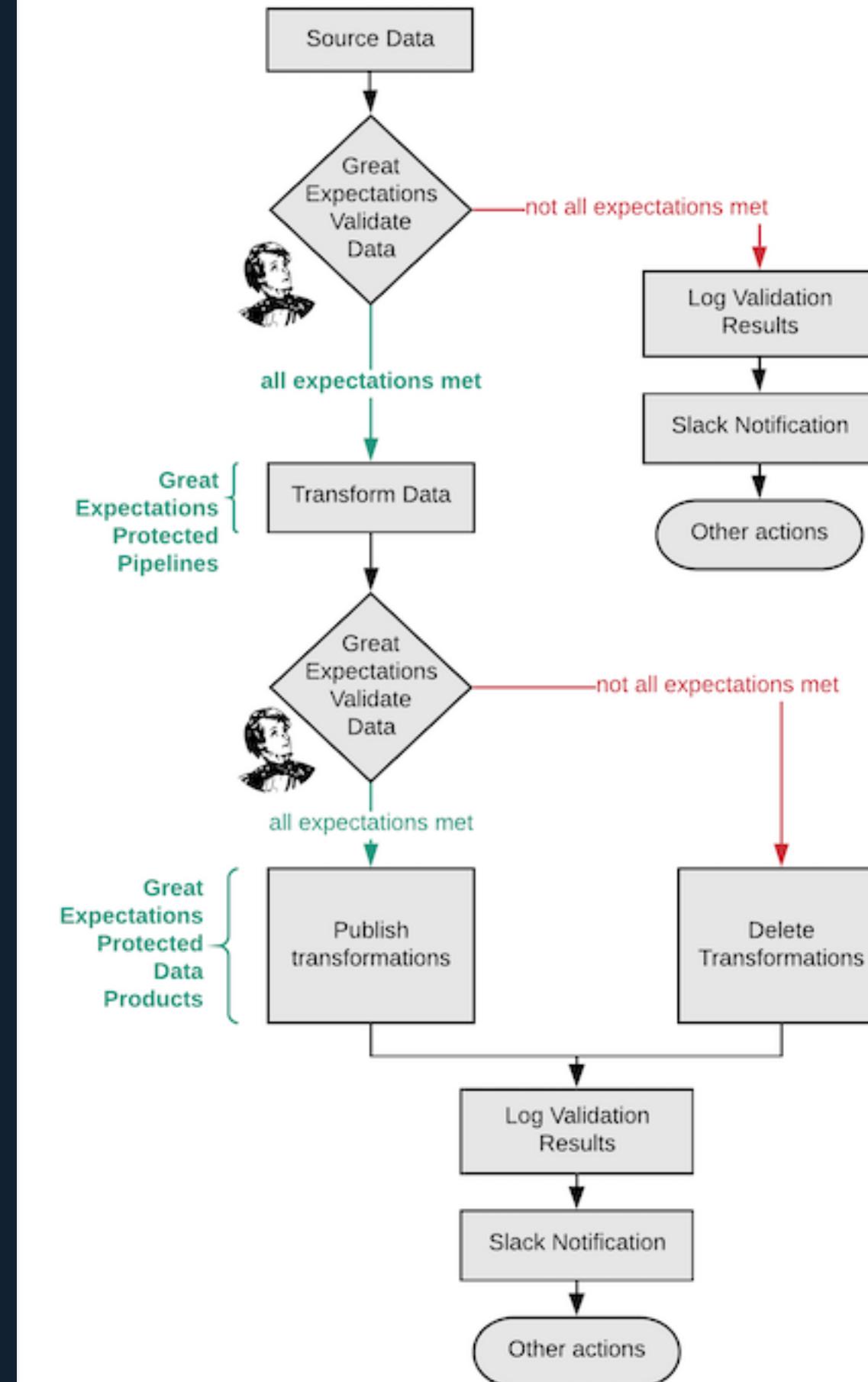
² Scaling Data Quality at Netflix <https://www.slideshare.net/MichelleUfford/scaling-data-quality-netflix-76917740>



HOW TO USE GREAT EXPECTATIONS IN A PIPELINE?

The WAP pattern!²

² Scaling Data Quality at Netflix <https://www.slideshare.net/MichelleUfford/scaling-data-quality-netflix-76917740>



QUESTION TIME



LIVE DEMO

QUESTION TIME



LET'S WORK TOGETHER!

OPEN SOURCE

Slack: Questions, ideas, random data banter

GitHub: Feature requests, bugs, pull requests

Deploy-a-thons: Work with us to get set up in a 1/2 day

OPTIONS WHERE MONEY CHANGES HANDS:

Open source support contracts

Lighthouse partnerships

THANK YOU!

Give us a look 🐞 greatexpectations.io

Give us a shout 🔈 greatexpectations.io/slack

Give us a star ⭐ github.com/great-expectations/great_expectations

Give us a try 🚀 `pip install great_expectations`

MISC BACKSTORY FUNDING & PHILOSOPHY

We raised a round from CRV and Root Ventures late last year, so we have two full years of runway to work on Great Expectations.

At some point, we'll build a paid SaaS product on top of Great Expectations, but for now we're just making the open source project as useful as it can be.

We're firmly committed to keeping Great Expectations open: everything that is open source will always remain open source.