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


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

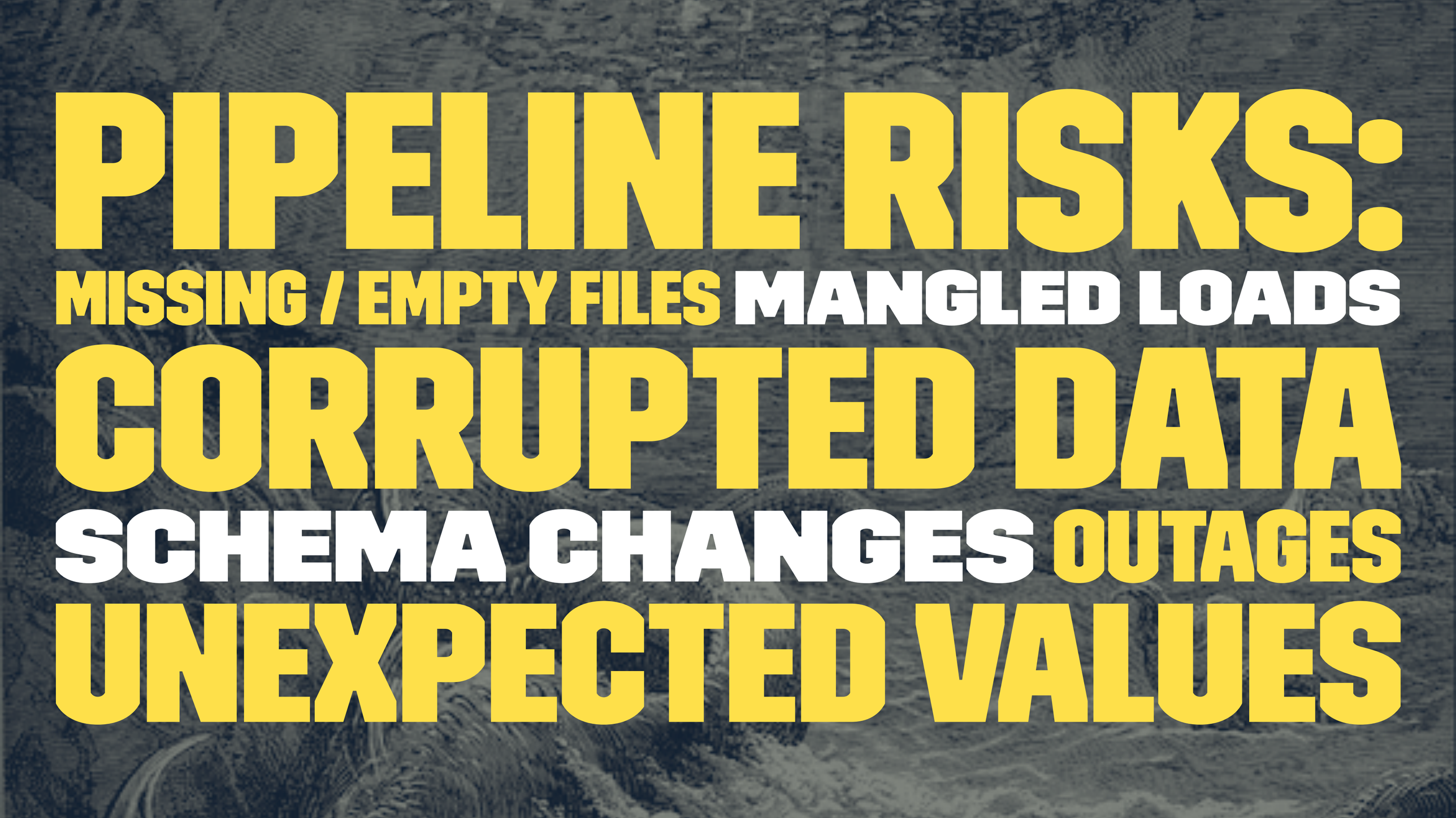
PIPELINE DEBT

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 more[1]

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

WHERE'S DOES THE COMPUTE HAPPEN?

GREAT EXPECTATIONS USES DIFFERENT BACK-END COMPUTE ENGINES

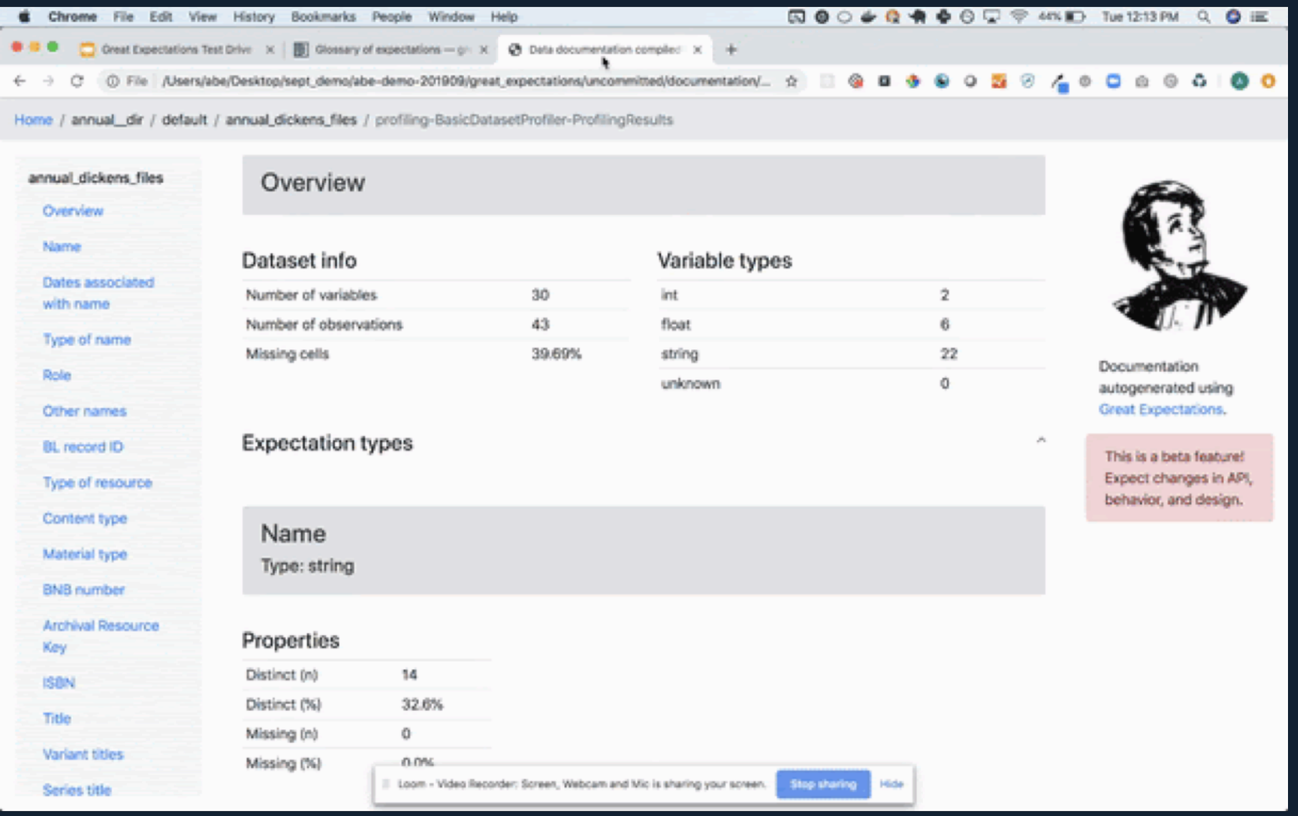
- pandas
- Spark
- Common relational databases via SQLAlchemy
  - Redshift
  - BigQuery
  - Snowflake
  - Postgres
  - MySQL

TESTS ARE DOCS

AND

DOCS ARE TESTS

TESTS ARE DOCS AND DOCS ARE TESTS



» 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 your team

BEFORE & AFTER ML MODELS

prevent malfeasant AI

ANALYTIC WAREHOUSES

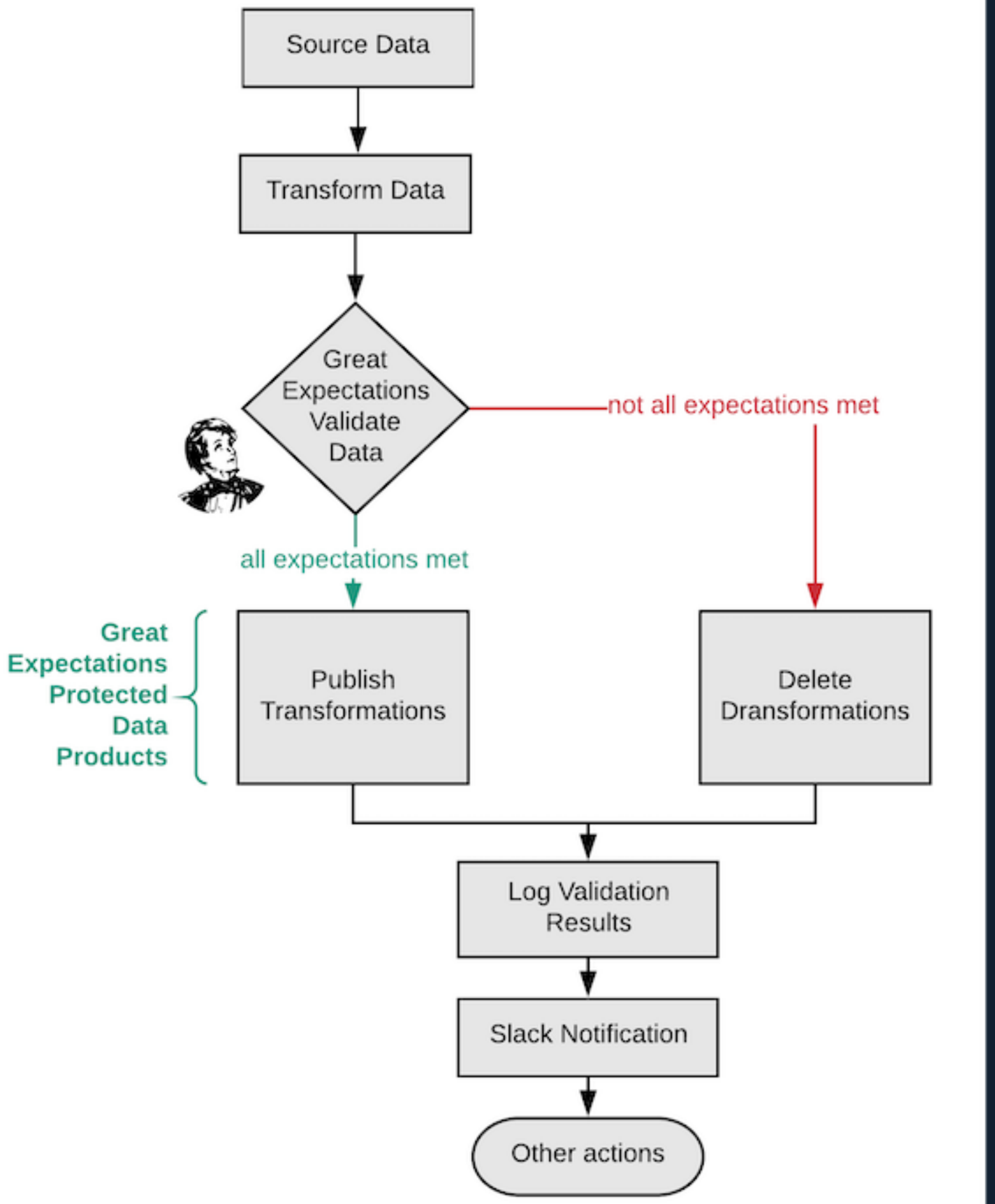
Be good data driven.

CRITICAL DASHBOARD TABLES

Don't piss off an executive

HOW TO USE GREAT EXPECTATIONS IN A PIPELINE?

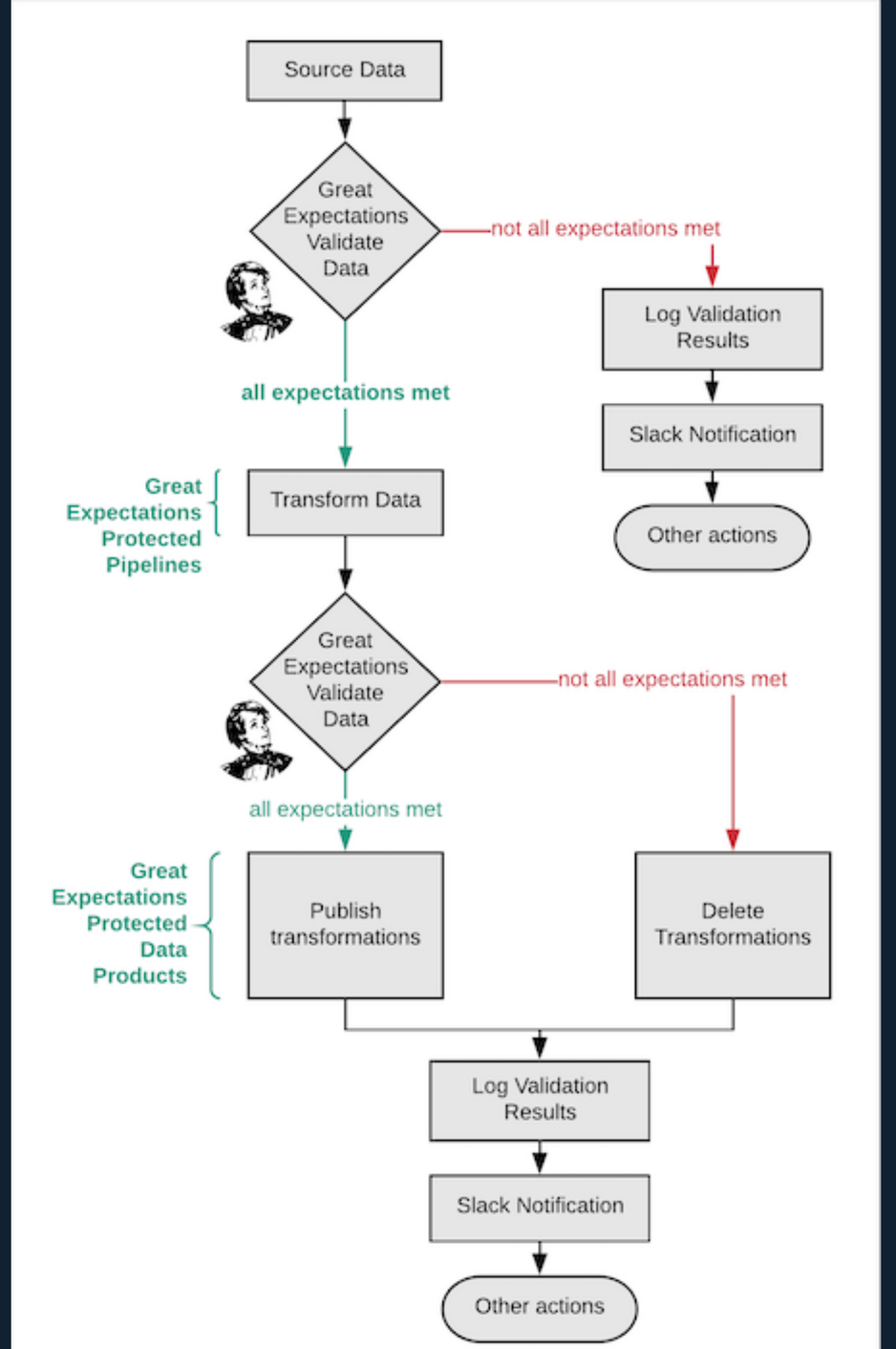
The WAP pattern![2]



[2] Scaling Data Quality at Netflix https://www.slideshare.net/MichelleFUfford/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/MichelleFUfford/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.