Testing and documenting your data doesn't have to suck

Data Council NYC - Nov 2019

great_expectations

SUPERCONDUCTIVE

SUPERCONDUCTIVE

@abeGong

About me (Abe)

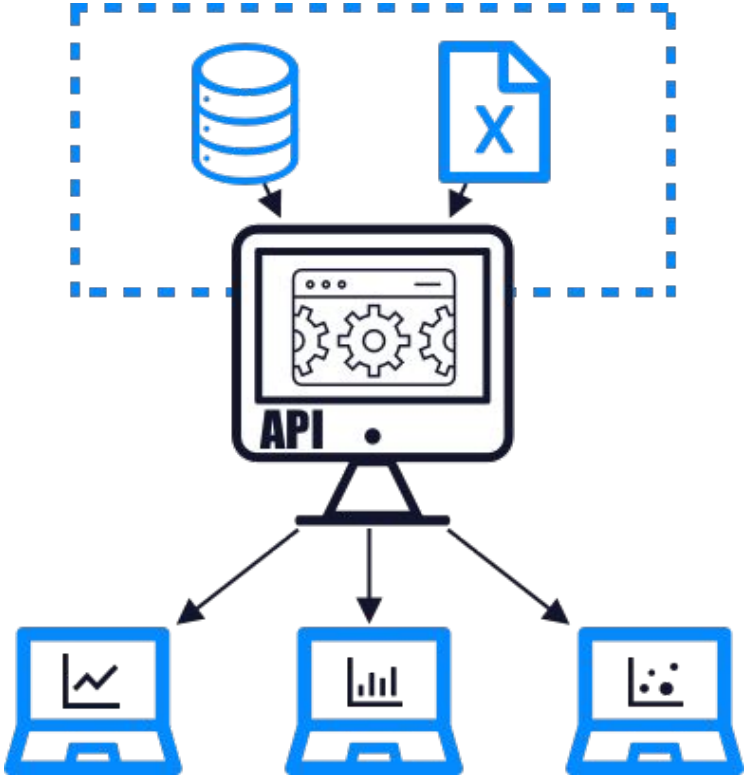
- Data scientist/engineer
- Tech-first and “enterprise”
- Human-scale, ethical data
- First time in NYC as an adult (?!)

Outline

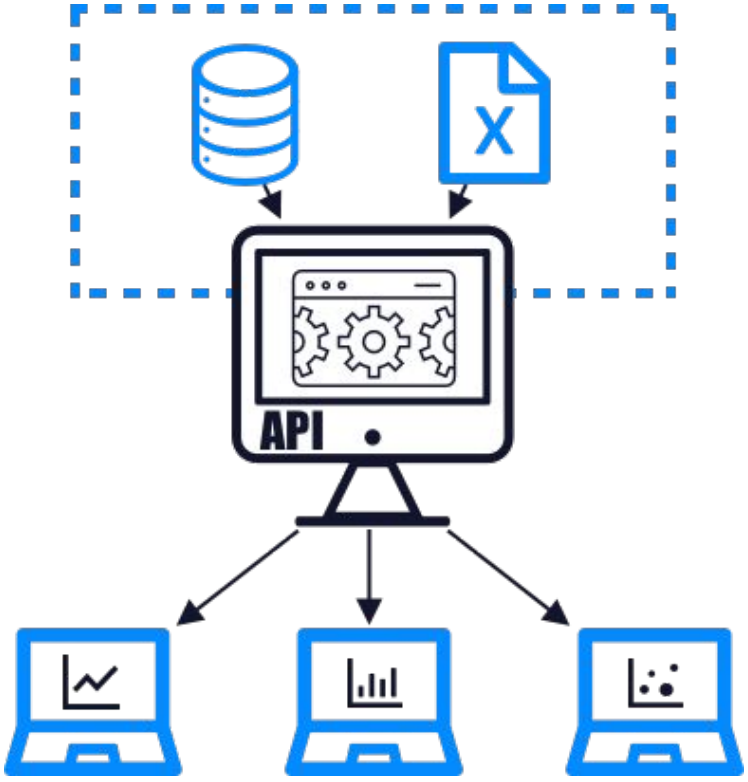
1. A thing we do that is ABSOLUTELY CRAZY
2. How to defeat pipeline debt
3. Volunteers wanted!


a thing we do
that is
ABSOLUTELY CRAZY
THIS IS MADNESS!

a thing we do that is ABSOLUTELY CRAZY

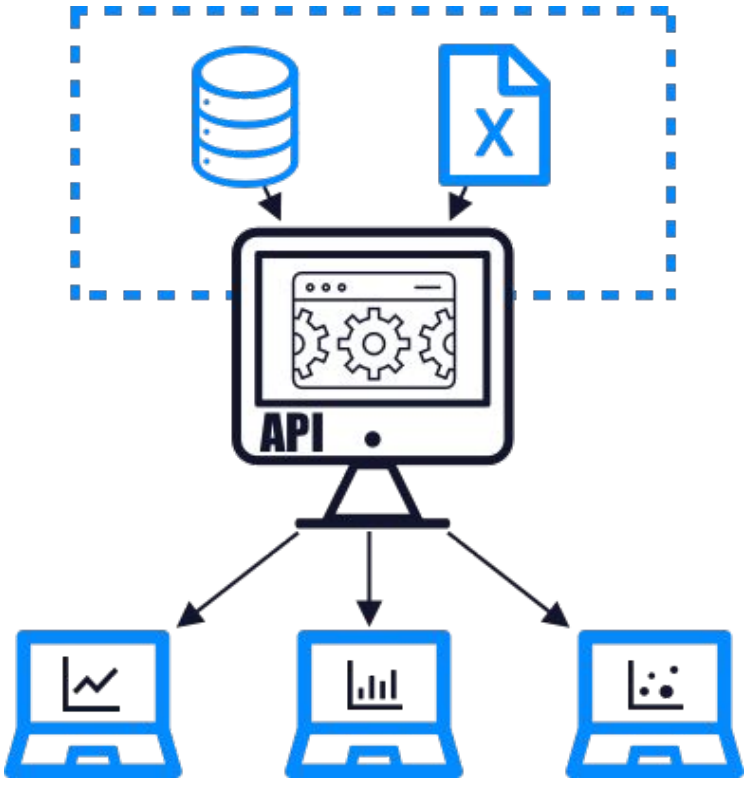


a thing we do that is ABSOLUTELY CRAZY



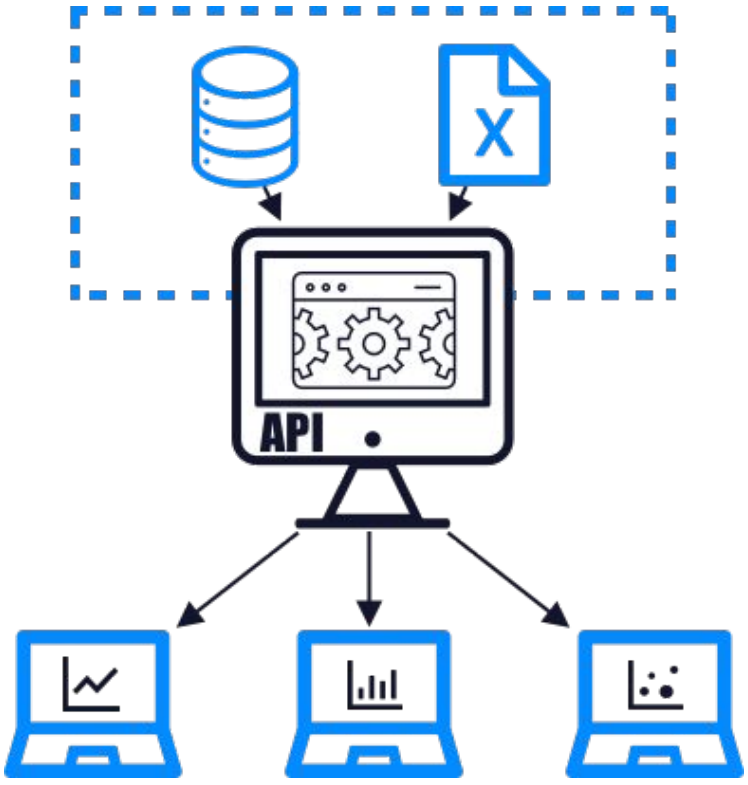
Undocumented

a thing we do that is ABSOLUTELY CRAZY



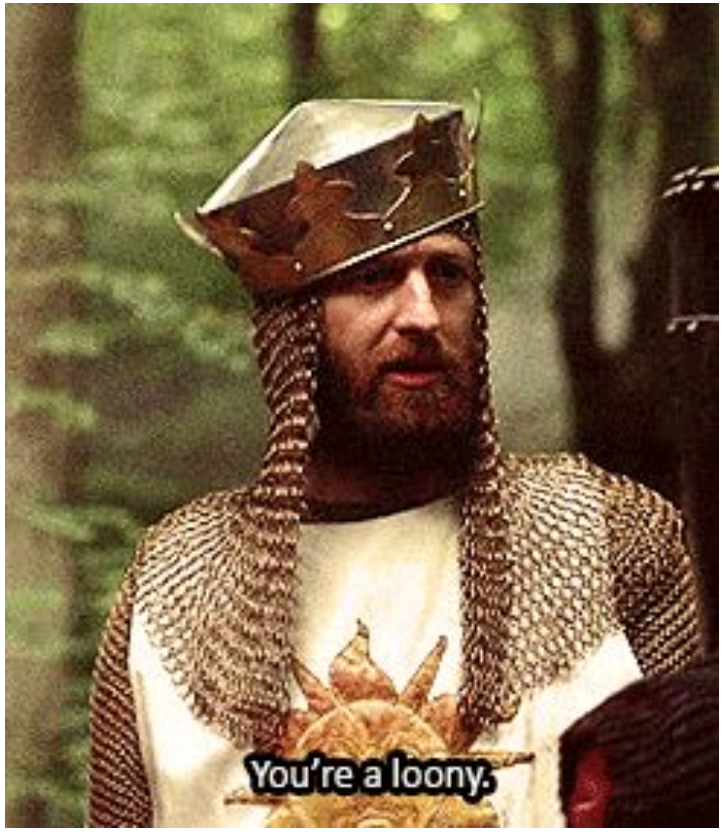
Undocumented
Untested

a thing we do that is ABSOLUTELY CRAZY



Undocumented
Untested
Unstable

a thing we do that is ABSOLUTELY CRAZY

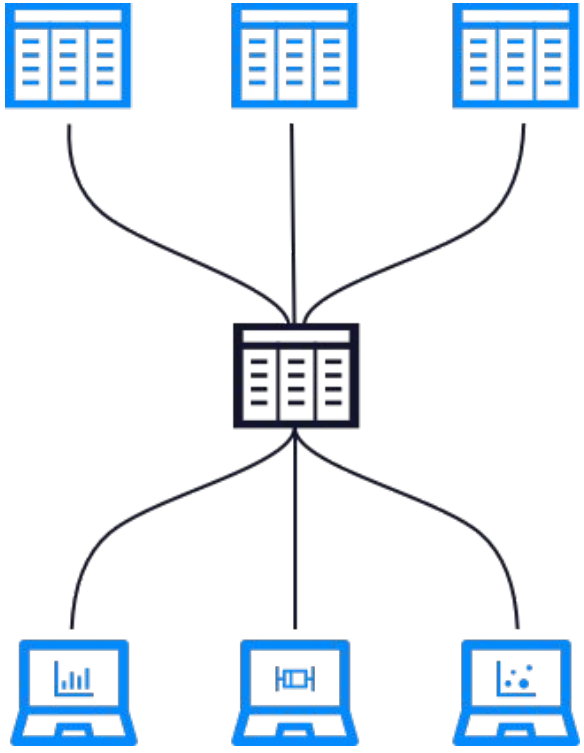


Undocumented

Untested

Unstable

a thing we do that is ABSOLUTELY CRAZY

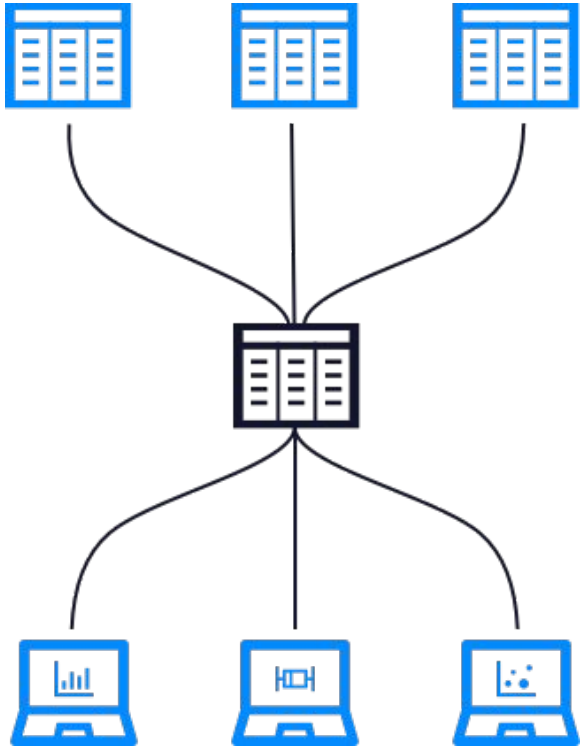


Undocumented

Untested

Unstable

a thing we do that is ABSOLUTELY CRAZY



Undocumented

Untested

Unstable

SUPERCONDUCTIVE

SUPERCONDUCTIVE

@abeGong

Trying to maintain a

data system

that is

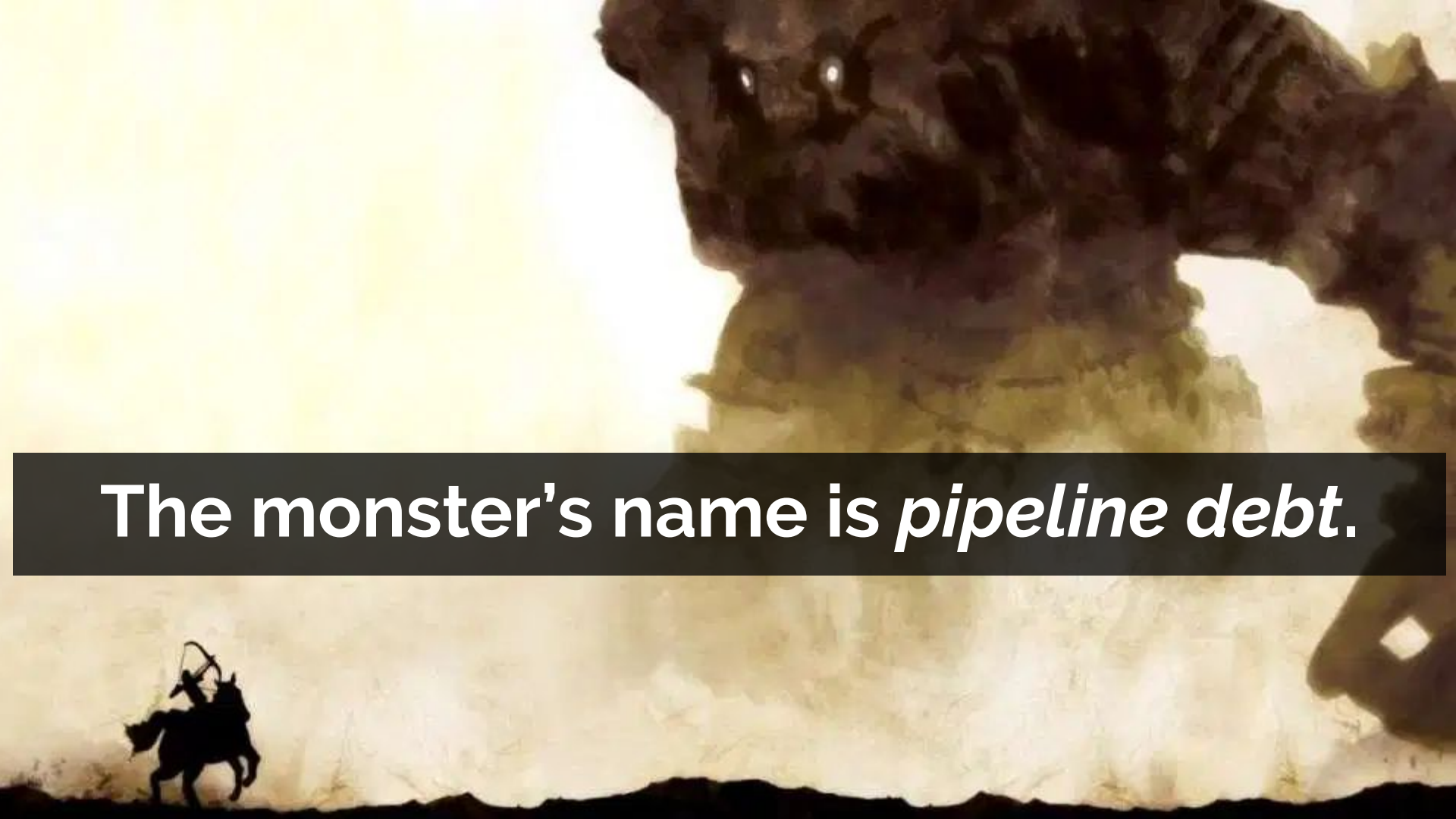
untested,

undocumented and

unstable

is

ABSOLUTELY CRAZY


The monster's name is *pipeline debt*.

great_expectations

Always know what to expect from your data

Expectations are assertions about data

great_expectations

```
expect_column_to_exist

expect_table_row_count_to_be_between
```

Expectations are assertions about data

great_expectations

```
expect_column_to_exist
expect_table_row_count_to_be_between
expect_column_values_to_be_unique
expect_column_values_to_not_be_null
expect_column_values_to_be_between
expect_column_values_to_match_regex
expect_column_values_to_match_strftime_format
```

Expectations are assertions about data

great_expectations

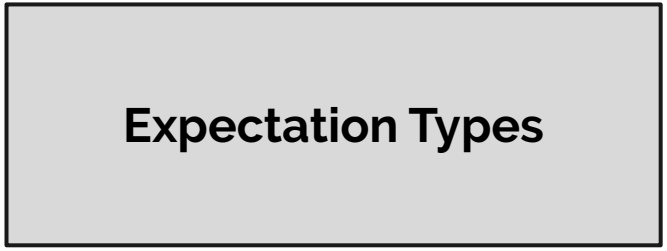
```
expect_column_to_exist
expect_table_row_count_to_be_between
expect_column_values_to_be_unique
expect_column_values_to_not_be_null
expect_column_values_to_be_between
expect_column_values_to_match_regex
expect_column_values_to_match_strftime_format
expect_column_mean_to_be_between
expect_column_kl_divergence_to_be_less_than
```

Expectations are assertions about data

great_expectations

```
expect_column_to_exist
expect_table_row_count_to_be_between
expect_column_values_to_be_unique
expect_column_values_to_not_be_null
expect_column_values_to_be_between
expect_column_values_to_match_regex
expect_column_values_to_match_strftime_format
expect_column_mean_to_be_between
expect_column_kl_divergence_to_be_less_than
etc. etc. etc.
```

Expectations are assertions about data



Expectations are assertions about data

Expectation Types

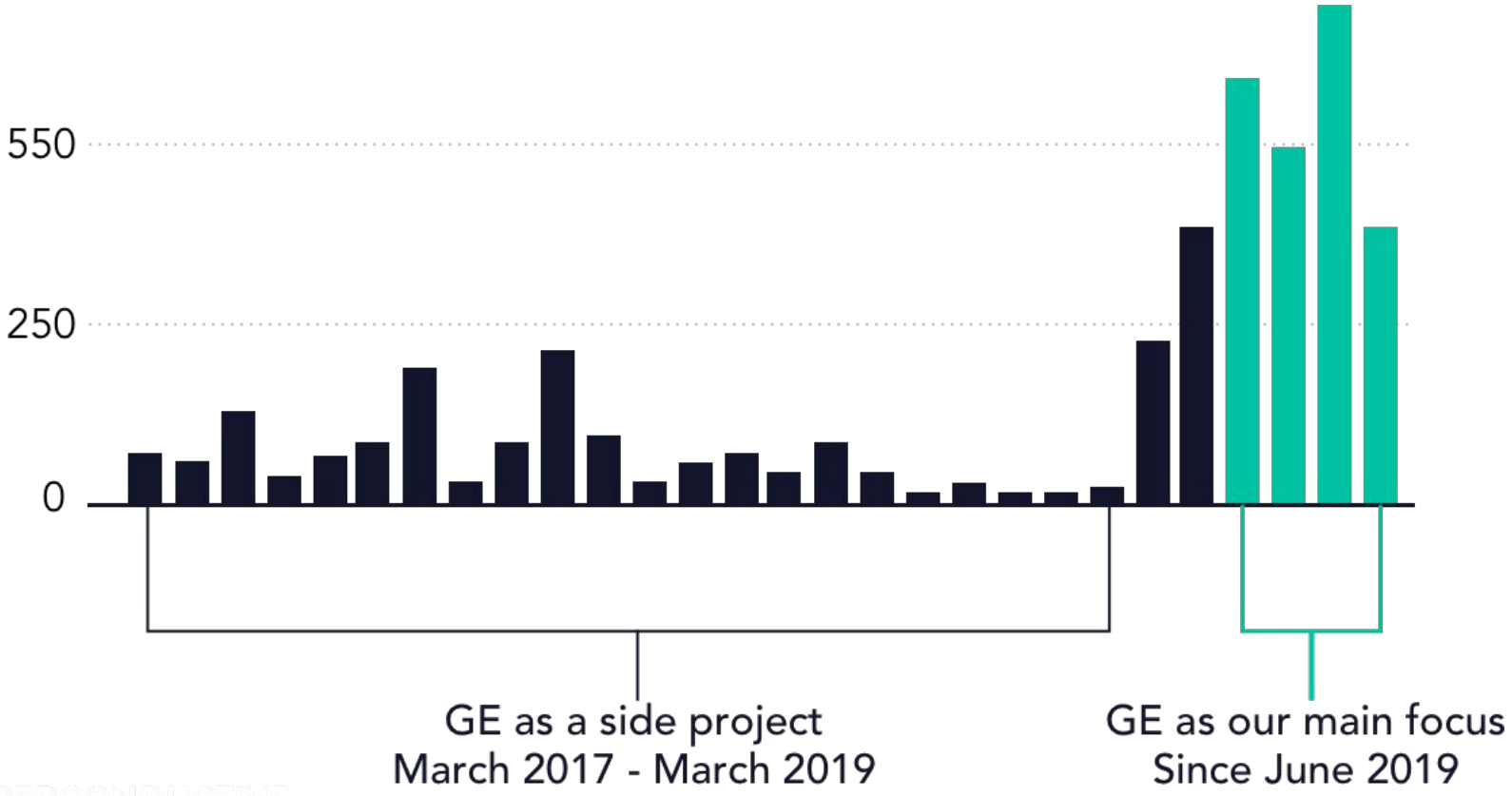
Data Sources

How to draw an owl

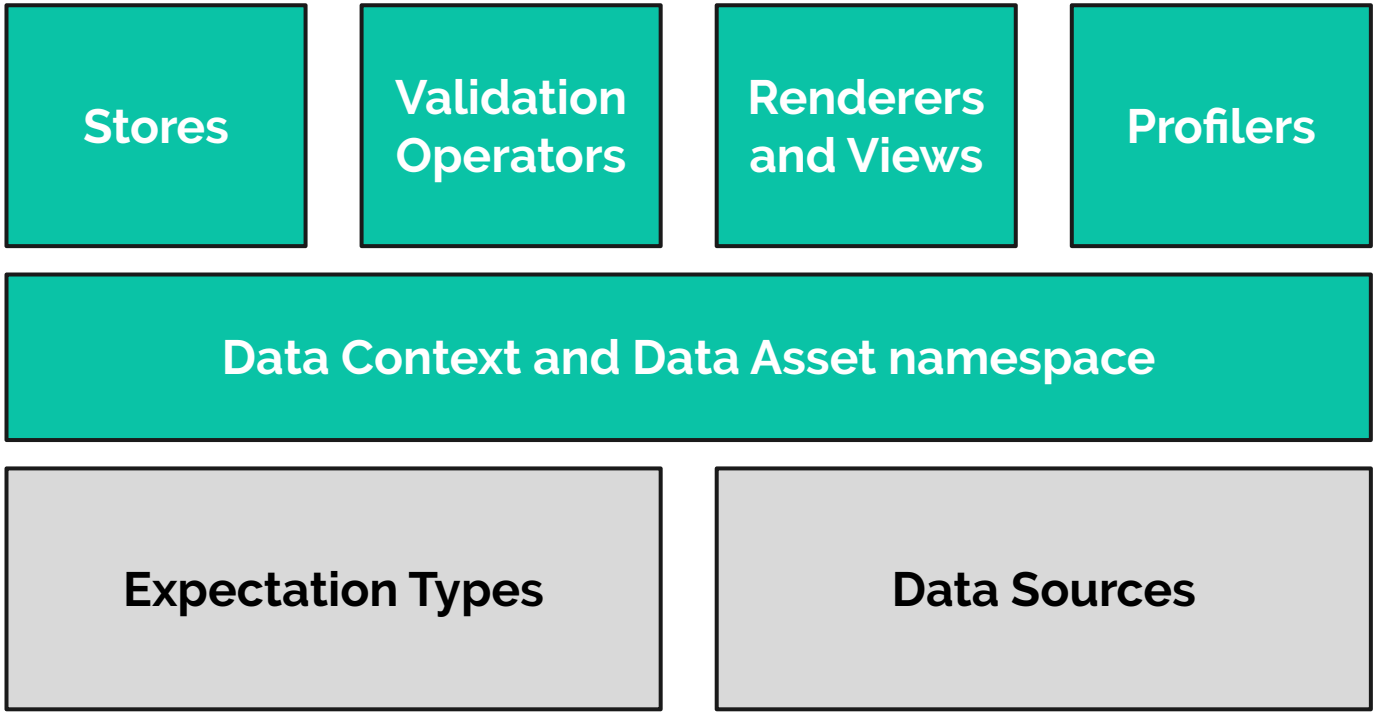
1. Draw some circles

2. Draw the rest of the stupid owl

Great Expectations has a bunch of shiny new features



Great Expectations has a bunch of shiny new features

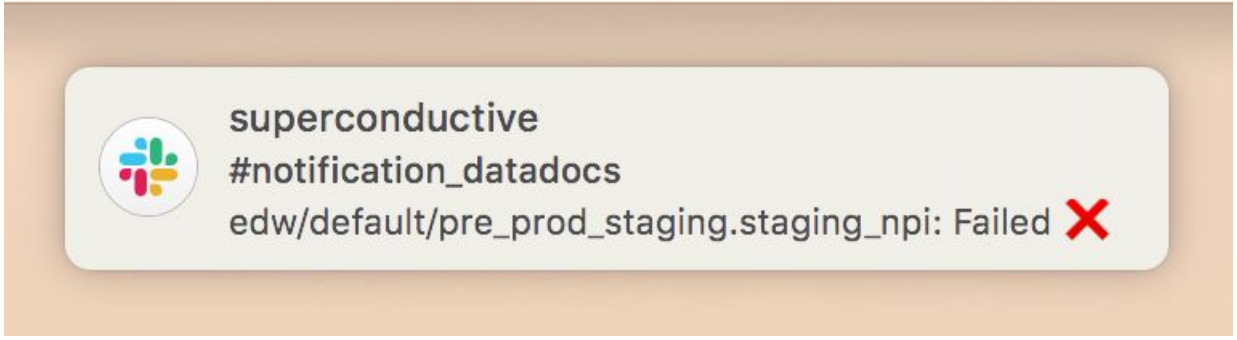


Great Expectations has a bunch of shiny new features

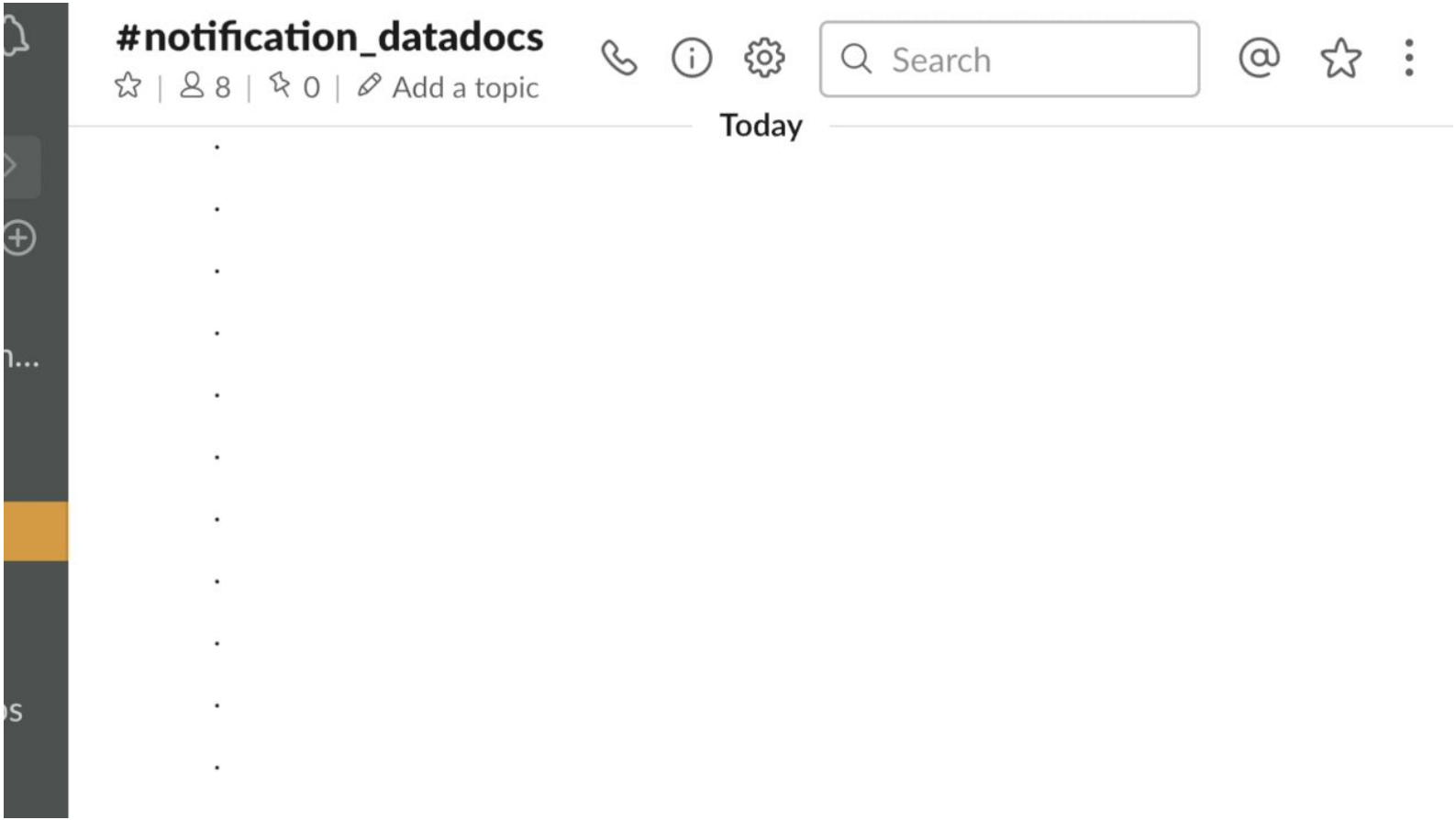
Great Expectations has a bunch of shiny new features

config_variables.yml

```
1  # This config file supports variable substitution which enables: 1) keeping
2  # secrets out of source control & 2) environment-based configuration changes
3  # such as staging vs prod.
4  validation_notification_slack_webhook: https://hooks.slack.com/services/d34db33f/
5
```



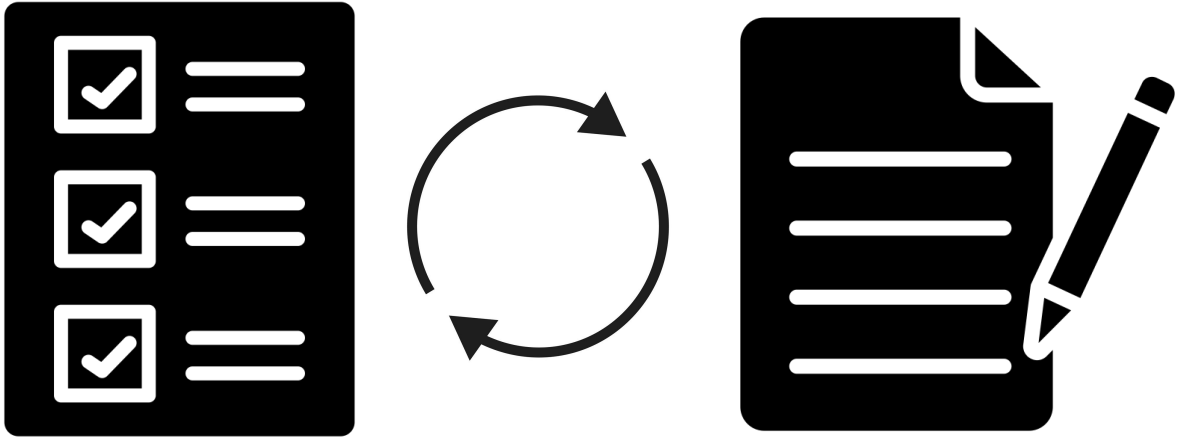
Great Expectations has a bunch of shiny new features



@abeGong

Set up data testing in a day, not a month.

Your docs are your tests,
and your tests are your docs.



Icons created by SBTS from Noun Project

Your docs are your tests, and your tests are your docs.

You probably don't need a data dictionary

ALEX JIA & MICHAEL KAMINSKY · JUN 17, 2019 · 10 MIN READ

While efforts to build a data dictionary are often undertaken out of a zeal for documentation that we would normally applaud, in practice data dictionaries and data catalogs end up being a large maintenance burden for little actual value, and tend to very quickly become out of date.

https://www.locallyoptimistic.com/post/data_dictionaries/

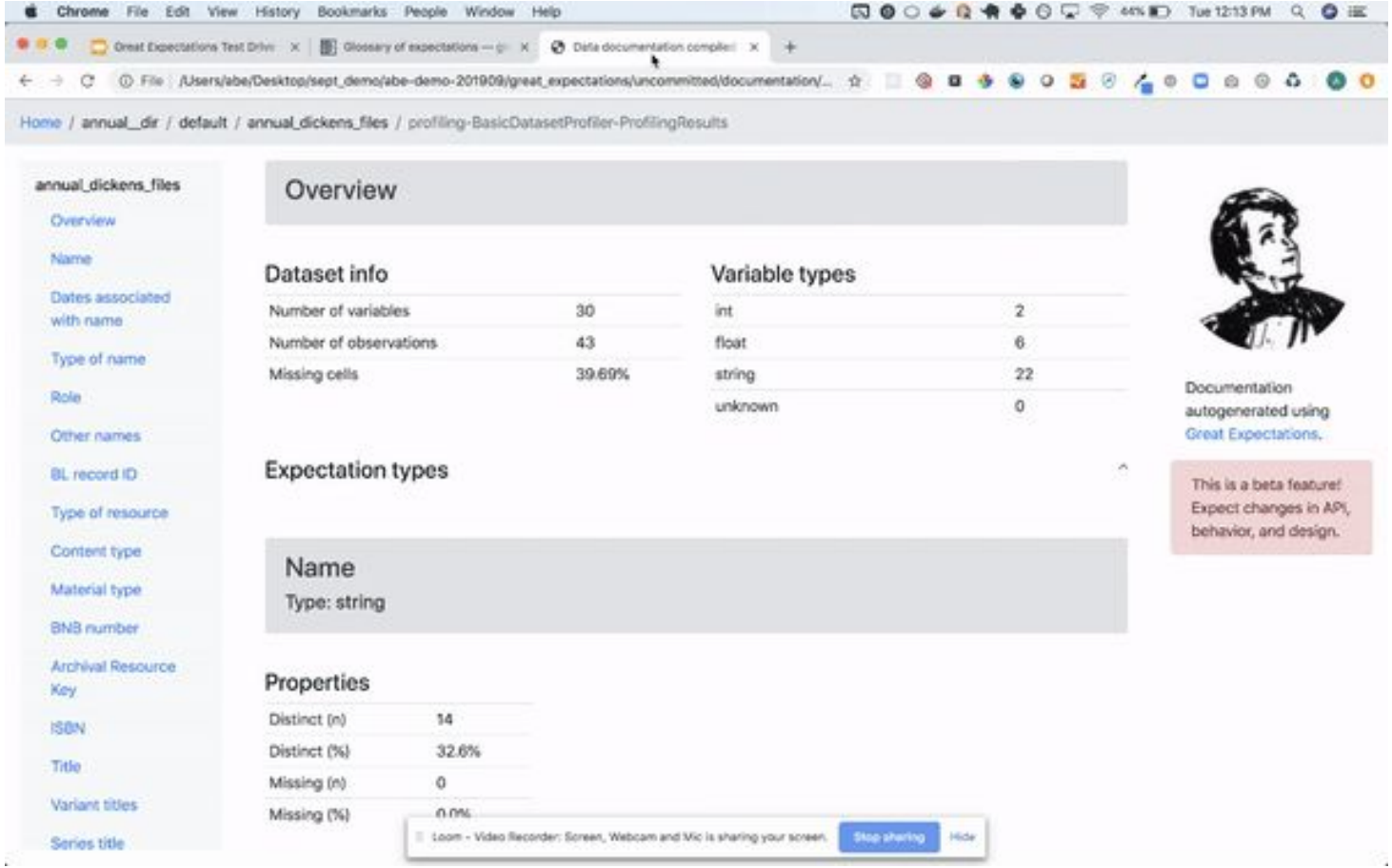
Your docs are your tests, and your tests are your docs.

```
expect_column_values_to_be_between(
    column="room_temp",
    min_value=60,
    max_value=75,
    mostly=.95
)
```


"Values in this column should be between 60 and 75, at least 95% of the time."

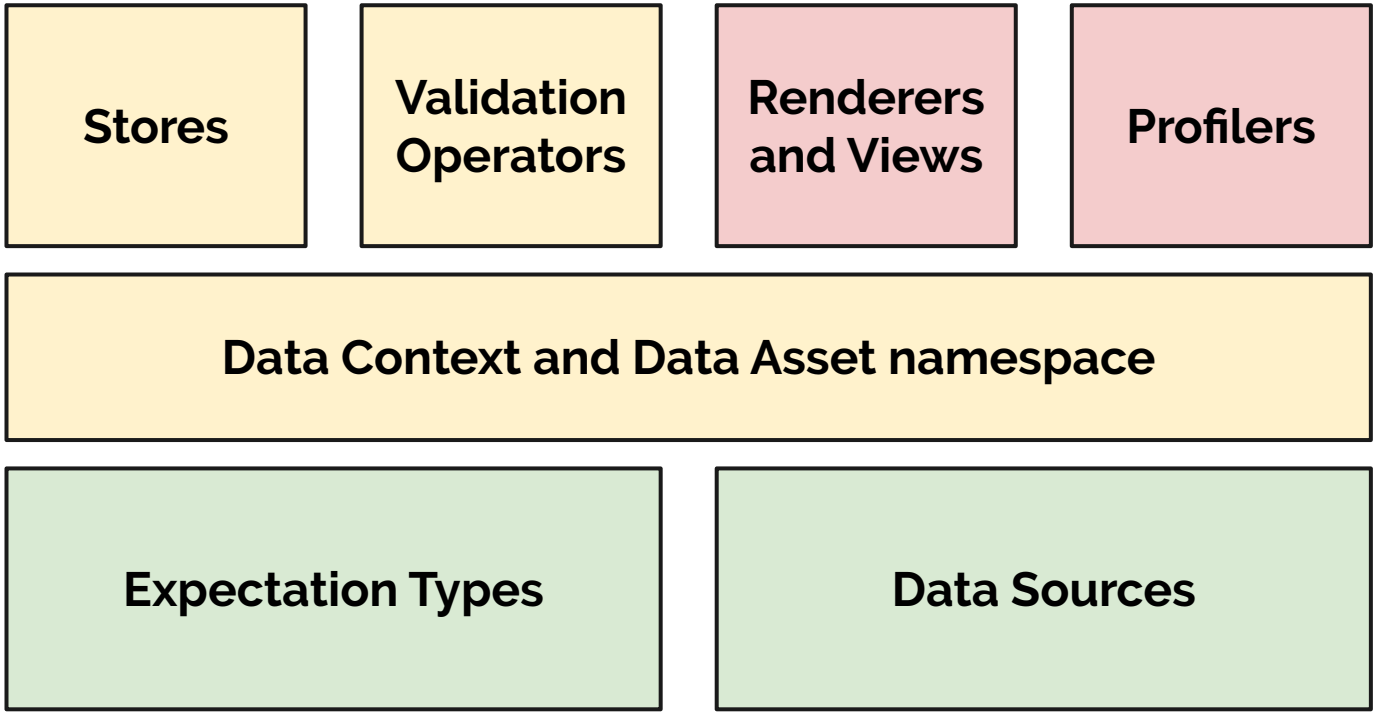
"Warning: more than 5% of values fell outside the specified range of 60 to 75."

Your docs are your tests, and your tests are your docs.



Warning: Great Expectations still has rough edges

Warning: Great Expectations still has rough edges



Volunteers wanted!

1. Pick a day
2. Work with us
3. Get set up
4. Improve the project

How to get in touch:

👋

https://greatexpectations.io/slack

Recap

Trying to maintain a

data system

that is

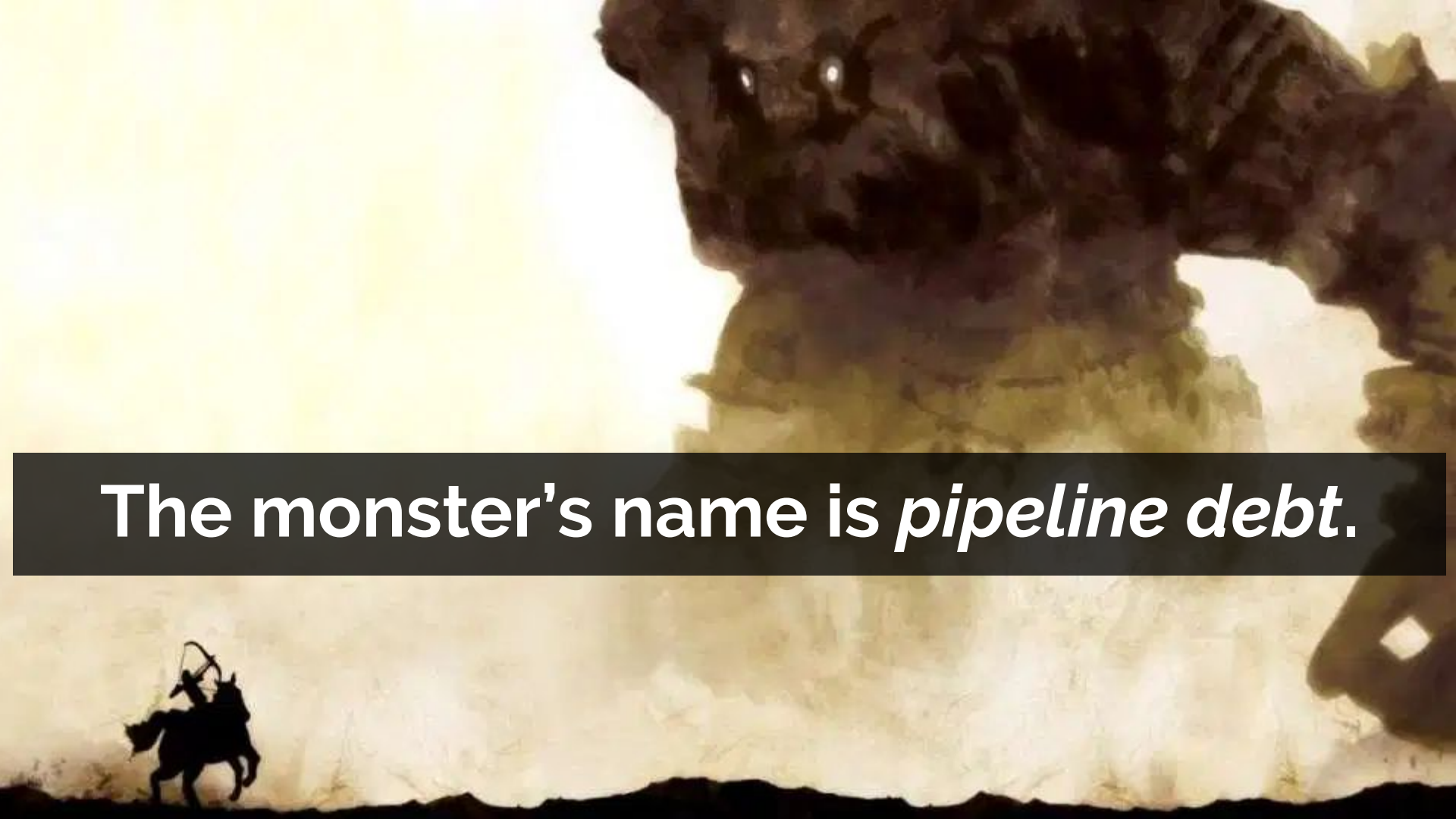
untested,

undocumented and

unstable

is

ABSOLUTELY CRAZY

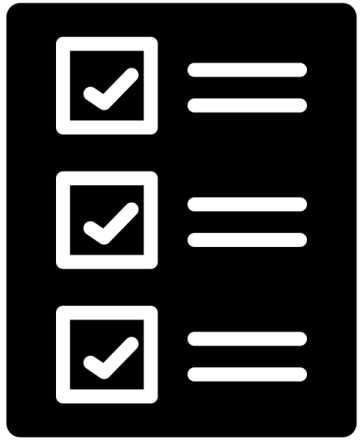

The monster's name is *pipeline debt*.

To defeat pipeline debt, always know what to expect of your data.

```
expect_column_to_exist
expect_table_row_count_to_be_between
expect_column_values_to_be_unique
expect_column_values_to_not_be_null
expect_column_values_to_be_between
expect_column_values_to_match_regex
expect_column_values_to_match_strftime_format
expect_column_mean_to_be_between
expect_column_kl_divergence_to_be_less_than
etc. etc. etc.
```

Set up data testing in a day, not a month.

Your docs are your tests,
and your tests are your docs.



Icons created by SBTS from Noun Project

Warning: Great Expectations still has rough edges

Volunteers wanted!

1. Pick a day
2. Work with us
3. Get set up
4. Improve the project

How to get in touch:

👋

https://greatexpectations.io/slack

Thank you, New York!

great_expectations

https://greatexpectations.io/slack