



Testing and documenting your data doesn't have to suck

Data Council NYC - Nov 2019

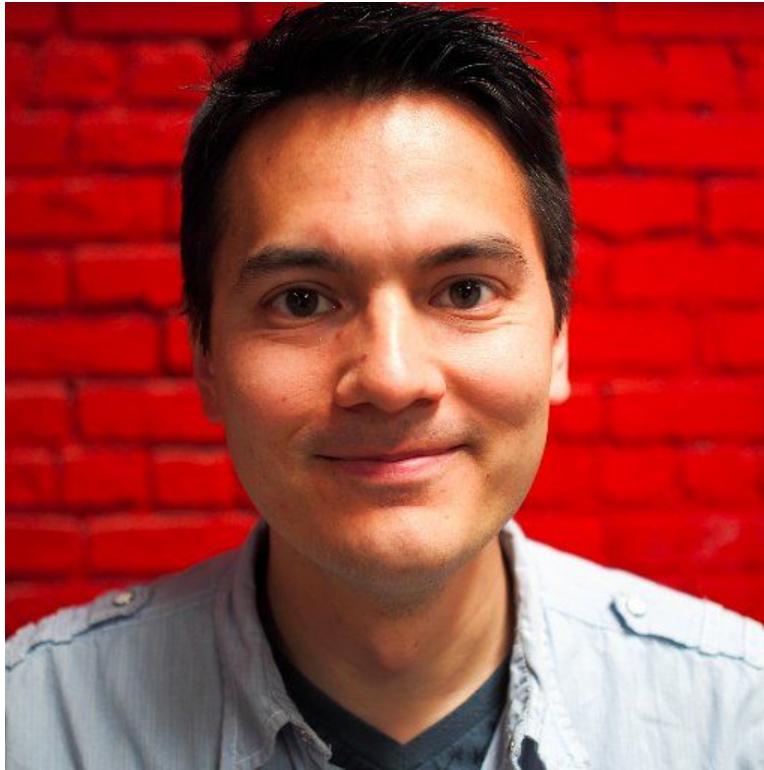


great_expectations



@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!



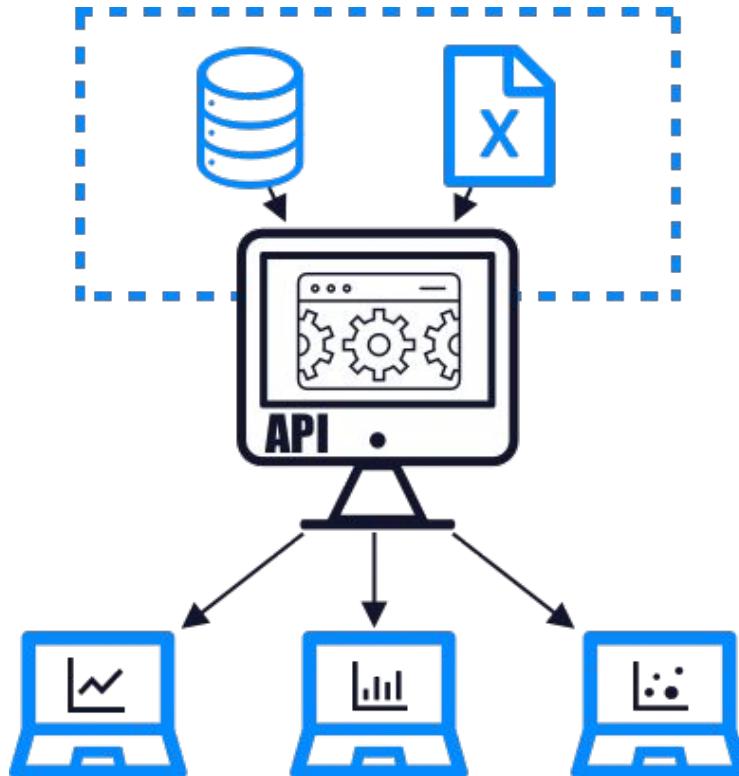
a thing we do

that is

ABSOLUTELY CRAZY

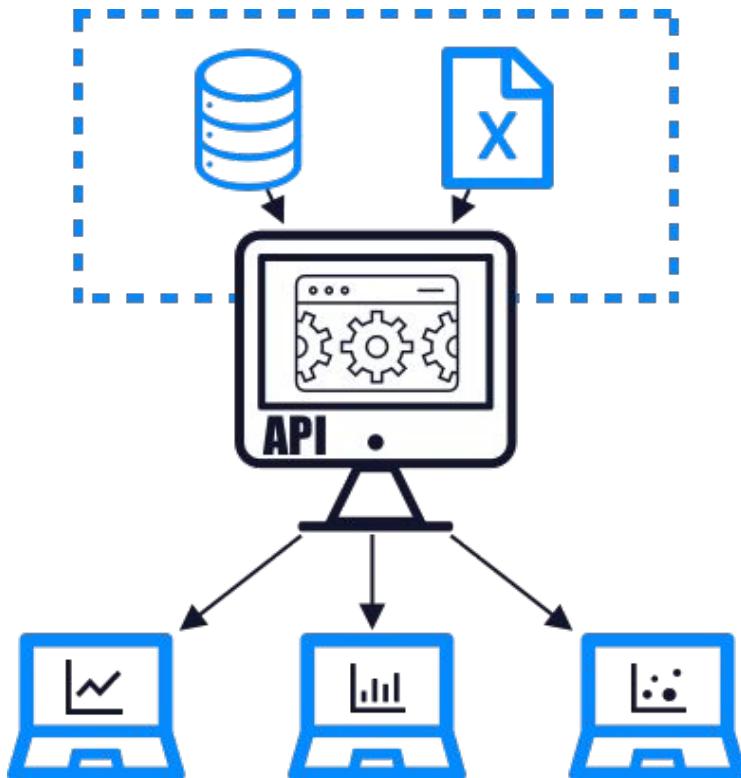
THIS IS MADNESS!

a thing we do that is ABSOLUTELY CRAZY

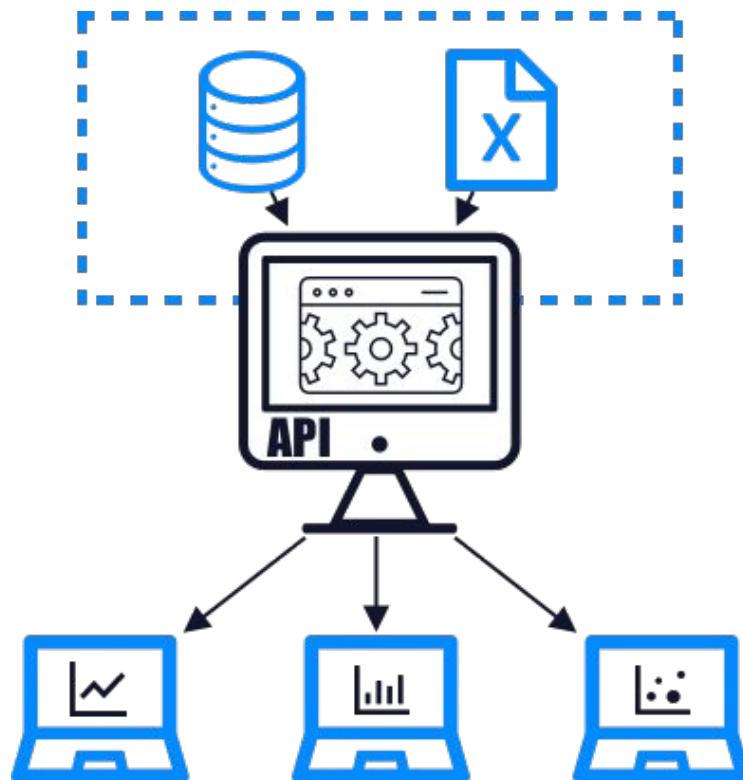


a thing we do that is ABSOLUTELY CRAZY

Undocumented

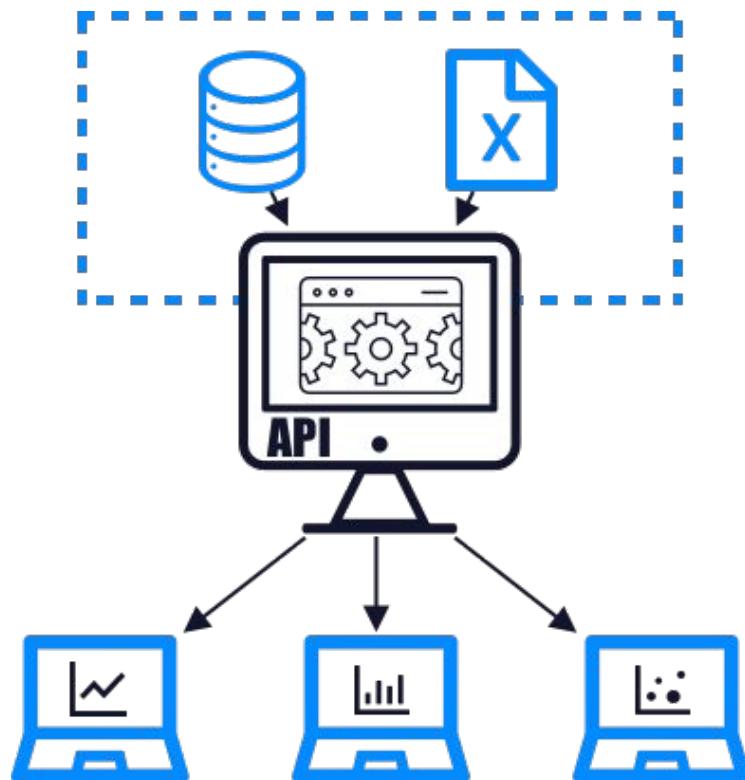


a thing we do that is ABSOLUTELY CRAZY



Undocumented Untested

a thing we do that is ABSOLUTELY CRAZY



Undocumented
Untested
Unstable

a thing we do that is ABSOLUTELY CRAZY



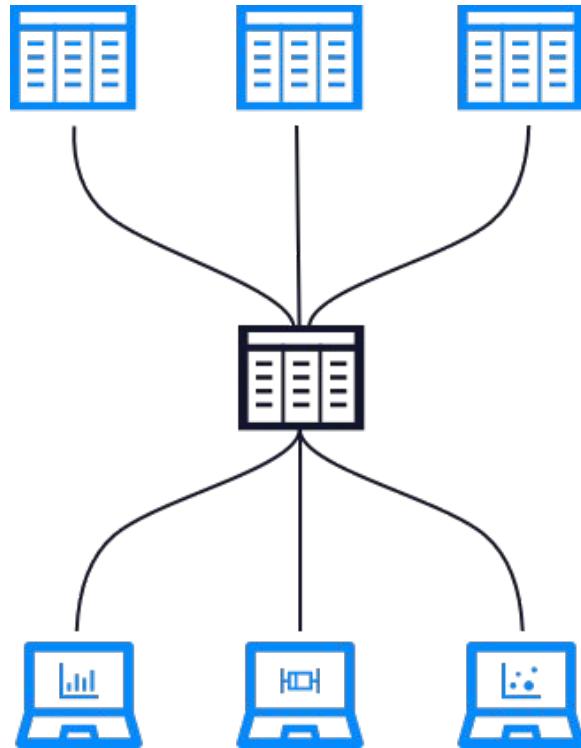
**Undocumented
Untested
Unstable**



SUPERCONDUCTIVE
SUPERCONDUCTIVE

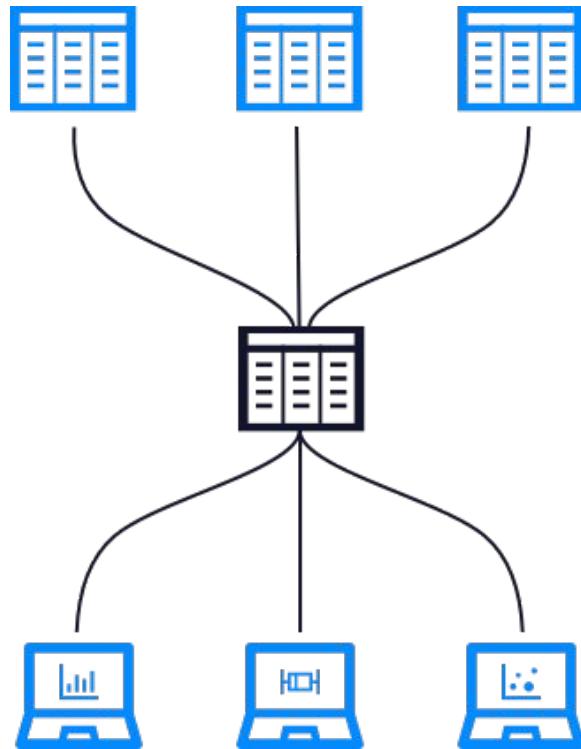
@abeGong

a thing we do that is ABSOLUTELY CRAZY



Undocumented
Untested
Unstable

a thing we do that is ABSOLUTELY CRAZY



**Undocumented
Untested
Unstable**



Trying to maintain a **data system**

that is

untested,

undocumented and

unstable

is

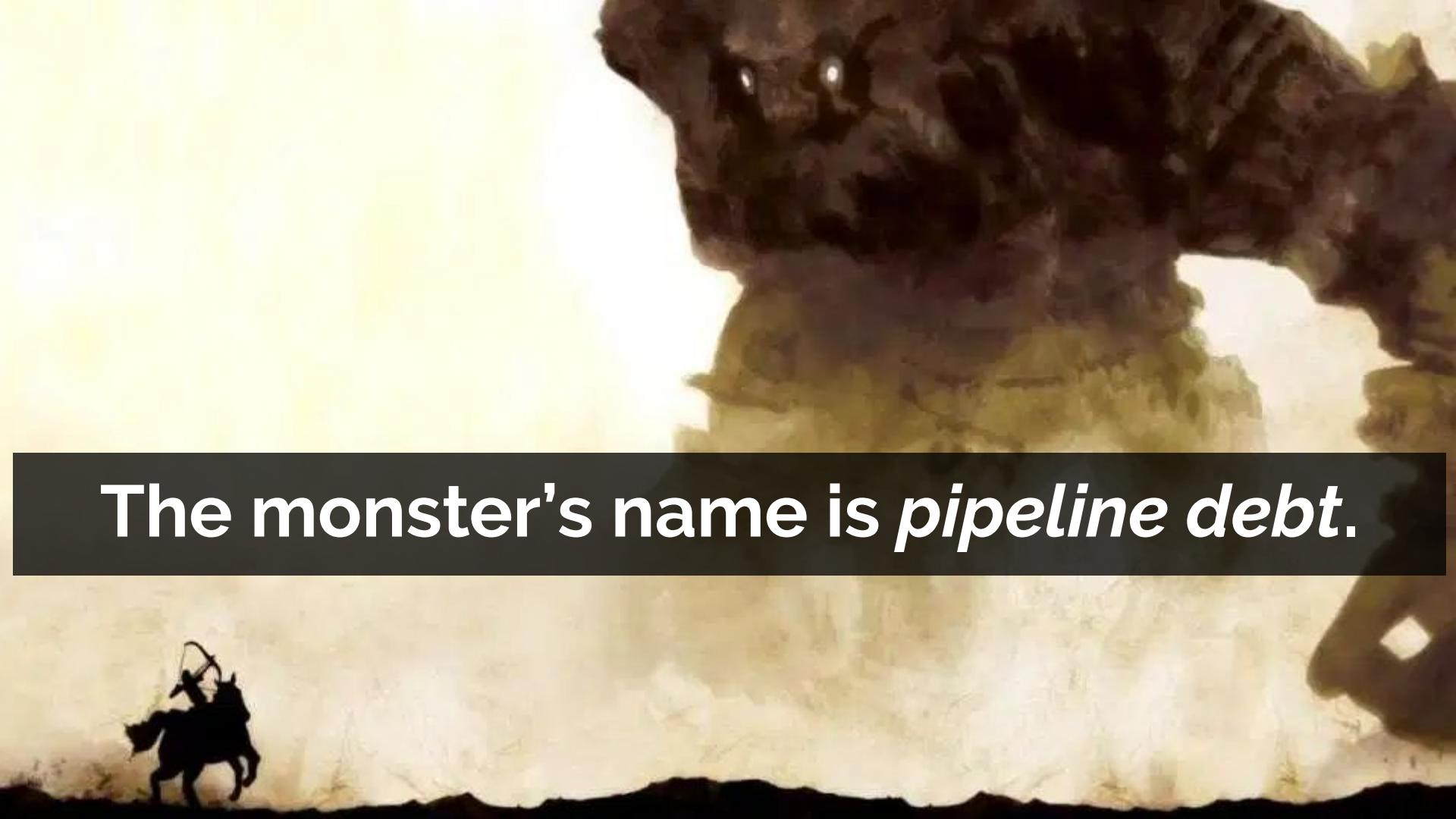
ABSOLUTELY CRAZY



?





A dramatic illustration featuring a large, dark, and smoky monster rising from a cloud of smoke. In the bottom left corner, a small silhouette of a person on a horse, holding a bow, looks up at the monster. The background is a bright, hazy yellow.

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

Expectation Types

Expectations are assertions about data

pandas

Spark

SQLAlchemy



Google
BigQuery

Expectation Types

Data Sources

How to draw an owl

1.



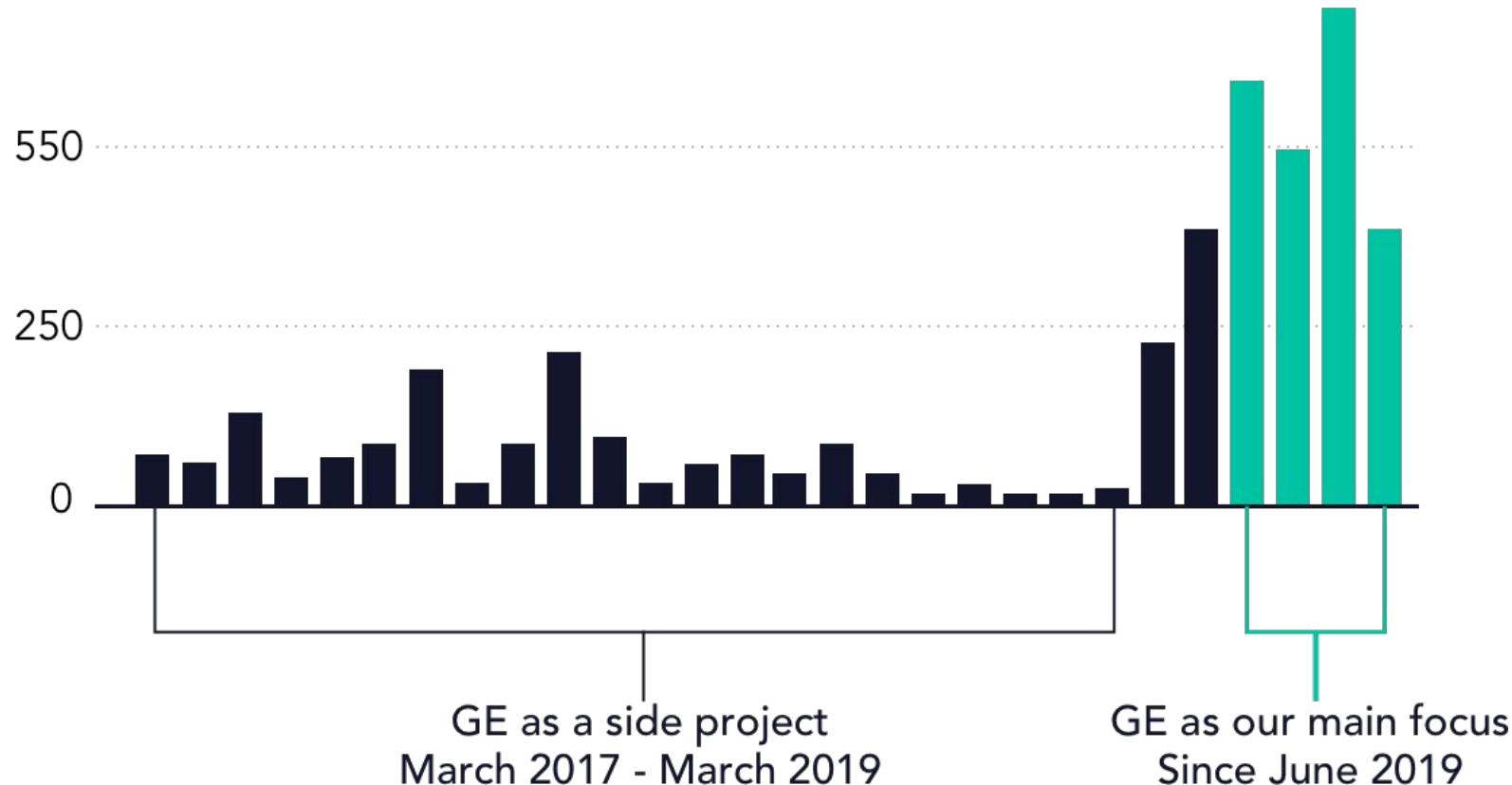
2.



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

Stores

Validation
Operators

Renderers
and Views

Profilers

Data Context and Data Asset namespace

Expectation Types

Data Sources

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
```



superconductive
#notification_datadocs
edw/default/pre_prod_staging.staging_npi: Failed ✘

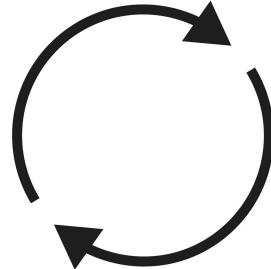
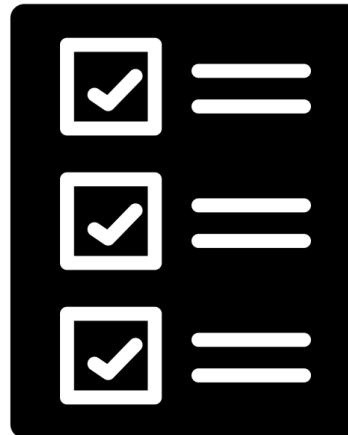
Great Expectations has a bunch of shiny new features

A screenshot of a dark-themed application interface. On the left is a vertical sidebar with a dark grey header containing a bell icon, a plus sign icon, and a truncated text 'n...'. Below this is a orange bar, followed by a dark grey footer with the letter 'S'. The main area has a light grey background. At the top, there's a header with the text '#notification_datadocs' in bold, followed by a star icon, a user count of '8', a comment count of '0', and a 'Add a topic' button. To the right of the header are icons for a phone, information, settings, and a search bar with the placeholder 'Search'. Below the search bar is a '@' icon, a star icon, and a three-dot menu icon. A horizontal line separates this from the main content area, which contains several small, faint dots indicating a list of items.

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



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



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.

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.

Screenshot of a web browser showing a dataset profile report for "annual_dickens_files".

The URL is: file:///Users/abe/Desktop/sept_demo/sbe-demo-201909/great_expectations/uncommitted/documentation/_profiling/BasicDatasetProfiler-Profilin

The page title is "annual_dickens_files - Overview".

Dataset info:

	Dataset info	Variable types
Number of variables	30	int
Number of observations	43	float
Missing cells	39.60%	string
		unknown
		0

Expectation types:

Name	Type: string

Properties:

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

A sidebar on the left lists various metadata fields:

- 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
- Series title

A small illustration of a person's head and shoulders is shown on the right.

A note states: "Documentation autogenerated using Great Expectations."

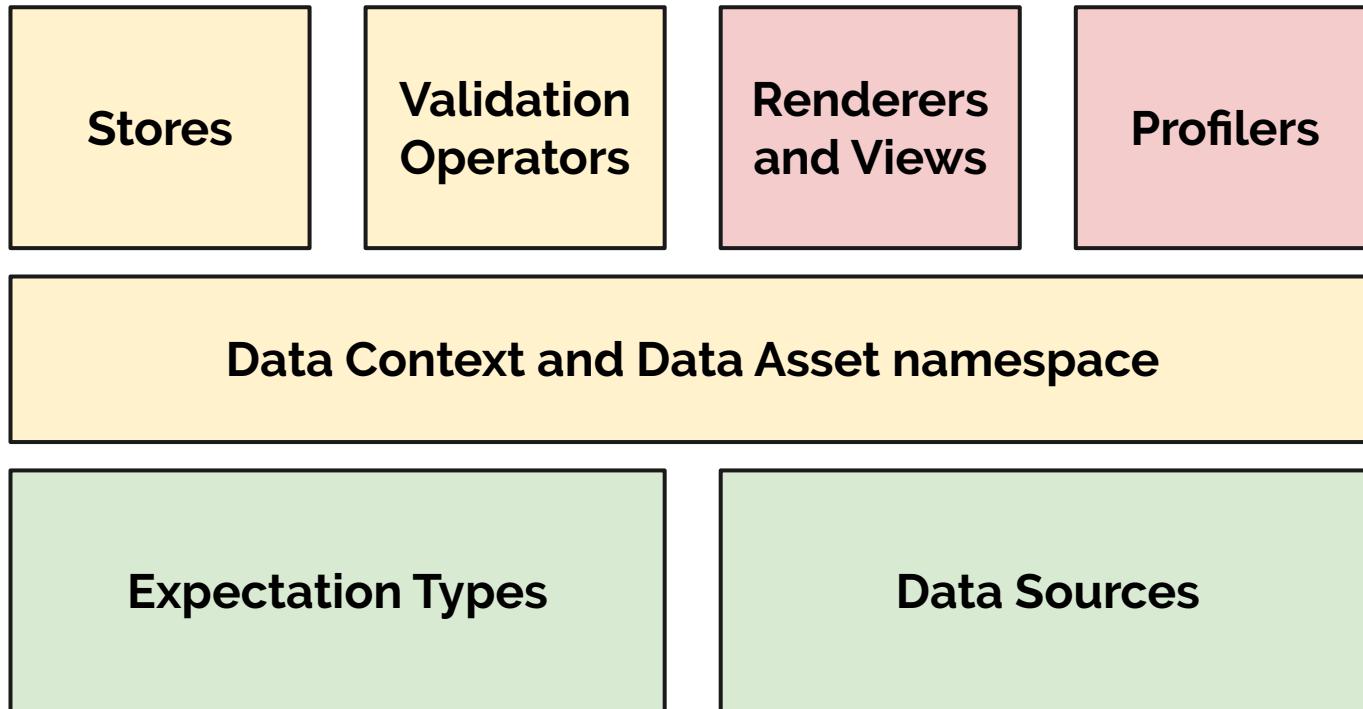
A red box contains the text: "This is a beta feature! Expect changes in API, behavior, and design."

Loom sharing information at the bottom: "Loom - Video Recorder: Screen, Webcam and Mic is sharing your screen." with "Stop sharing" and "Hide" buttons.

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>



imgflip.com

SUPERCONDUCTIVE

@abeGong

Recap

Trying to maintain a **data system**

that is

untested,

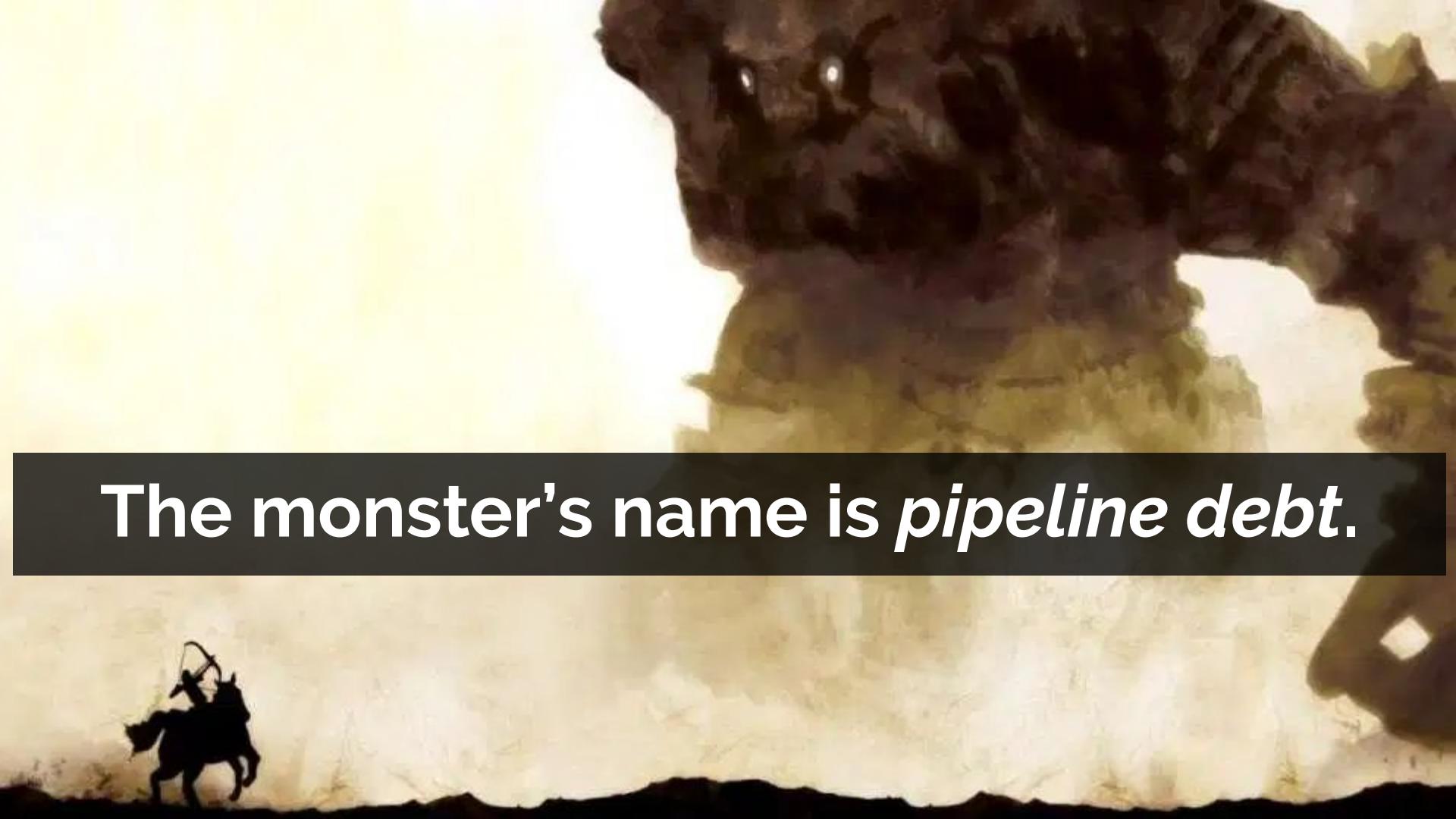
undocumented and

unstable

is

ABSOLUTELY CRAZY



A dramatic illustration featuring a large, dark, and smoky monster rising from the ground. In the bottom left corner, a small silhouette of a person on a horse, holding a bow, looks up at the monster. The background is a bright, hazy yellow and white.

The monster's name is *pipeline debt*.

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

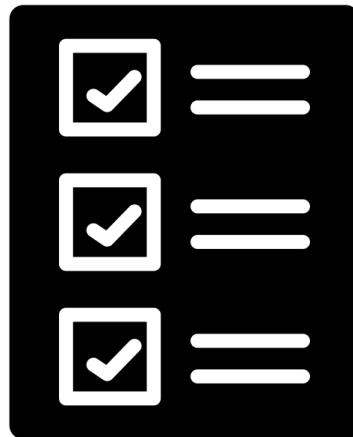


```
expect_column_to_exist  
expect_table_row_count_to_be_between  
expect_column_values_to_be_unique  
expect_column_values_to_not_be_null  
expect_column_values_to_be_between  
expect_column_values_to_match_regex  
expect_column_values_to_match_strftime_format  
expect_column_mean_to_be_between  
expect_column_kl_divergence_to_be_less_than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

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



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



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>



imgflip.com

SUPERCONDUCTIVE

@abeGong

Thank you, New York!



great_expectations

<https://greatexpectations.io/slack>