Test All The Data!

Eric J. Ma (MIT)

About Me

- PhD Candidate, MIT Biological Engineering.

About Me

- PhD Candidate, MIT Biological Engineering.
- I play with statistics and biological data.

About Me

- PhD Candidate, MIT Biological Engineering.
- I play with statistics and biological data.
- I think you need to write tests for your data.

Why?

1. You have data.

Why?

1. You have data.
2. You have assumptions of your data.

Why?

1. You have data.
2. You have assumptions of your data.
3. You will modify your data, making more assumptions about that data.

Why?

1. You have data.
2. You have assumptions of your data.
3. You will modify your data, making more assumptions about that data.

The data do not always follow your assumptions!

Example

- Data: one column needs to be log10-transformed.
- Assumptions?
  - `dtype`?
  - `range`?

HOW?!

Step 1: Install `py.test`

```
$ pip install pytest
```

Step 2: Create your test script.

```
$ touch test_data.py
$ nano test_data.py
```

Step 3: Make your script read data.

```
import pandas as pd
data = pd.read_csv('data.csv')
```

Step 4: Write your test functions.

```
def test_column_is_correct():
    assert data['column'].dtype == float
    assert data['column'].min() > 0

def test_data_state_integrity():
    assert 'log10_col' not in data.columns
```

Step 5: Run your tests.

```
$ py.test
```

Step 6: Scream at your data provider.

- Actually, talk nicely :)

Step 6: Scream at your data provider.

- Actually, talk nicely :)
- Discuss why data test failed.

Step 6: Scream at your data provider.

- Actually, talk nicely :)
- Discuss why data test failed.
- Fix data.

Step 6: Scream at your data provider.

- Actually, talk nicely :)
- Discuss why data test failed.
- Fix data.
- Test again.

Step 6: Scream at your data provider.

- Actually, talk nicely :)
- Discuss why data test failed.
- Fix data.
- Test again.

Rinse and repeat.

Step 6: Scream at your data provider.

- Actually, talk nicely :)
- Discuss why data test failed.
- Fix data.
- Test again.

Rinse and repeat.
You'll never have enough data tests!

Benefits

- An automated contract between yourself, your future self, and others on your team.

Benefits

- An automated contract between yourself, your future self, and others on your team.
- Encodes your sanity checks on the data, so you don't have to remember them individually the future.

Benefits

- An automated contract between yourself, your future self, and others on your team.
- Encodes your sanity checks on the data, so you don't have to remember them individually the future.
- If data changes, you have a check for integrity.

Conclusion

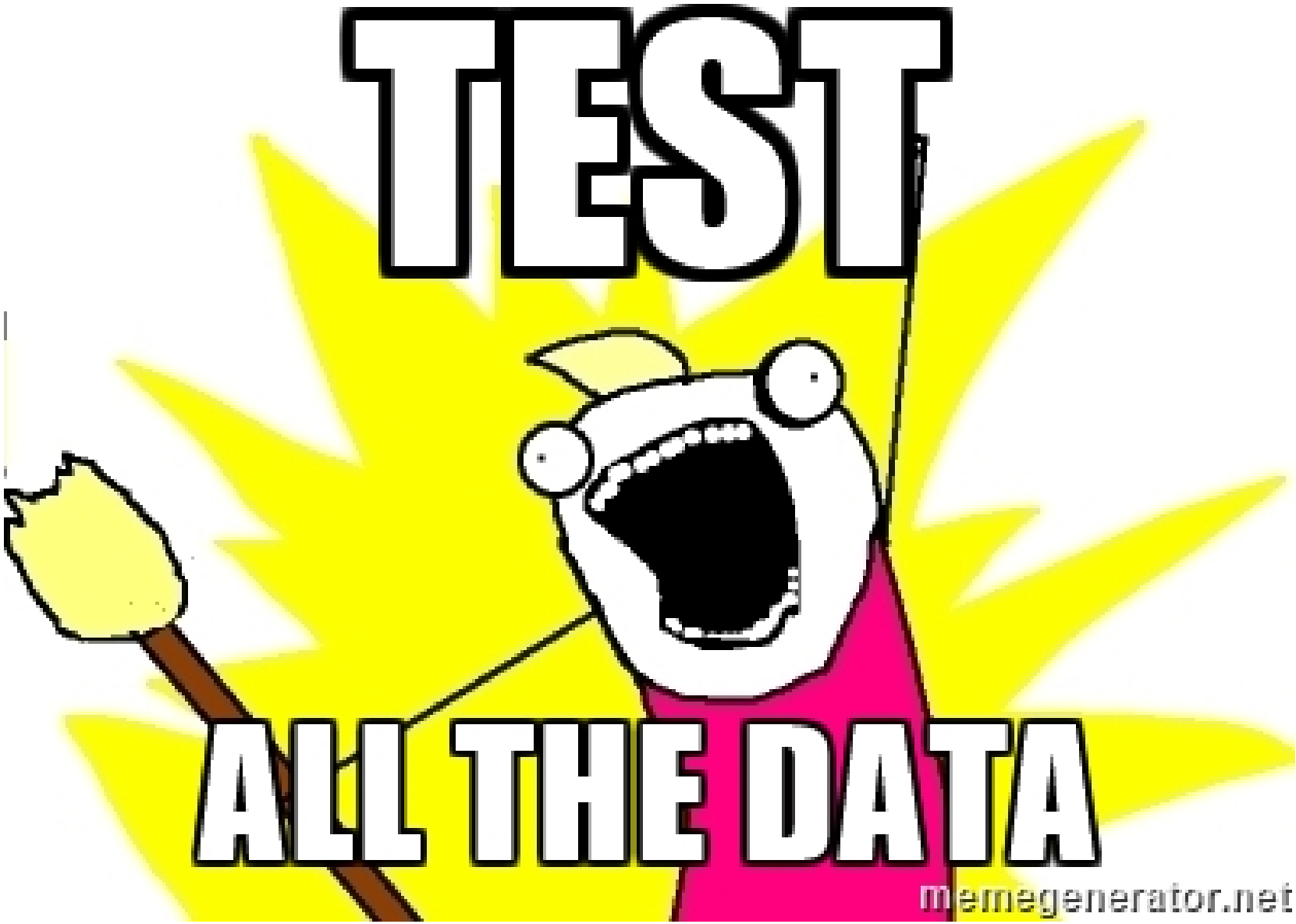

Figure 1: