SODA

On the importance of using a data quality framework to monitor your data.

Don’t Let Your Models Decay!

Bastien Boutonnet,
Lead Data Scientist

Bastien joined Soda last year and before that he was at TripActions and Travelbird, once he decided that the Postdoctoral Fellow wasn't half the fun ;-). He's a die hard dbt fan, DJ, and French person living in Amsterdam.

https://www.bastienboutonnet.com

© Soda Data NV. Proprietary and Confidential.

Zillow: A Cautionary Tale

$500m

Zillow **overestimated** the value of the houses it purchased in Q3 and Q4 of 2021 **by over $500m**

~25%

Q3 **losses of $304m**, leading into a **~25%** workforce reduction

Coincided with a **strong change in housing market conditions** which causes housing prices to fall

Strong evidence that their models were trained on the, then “old” situation which indicated growing prices, which caused their **unattended models to work under a different “assumption”** or concept

Did they do everything badly?

No!

- Their models were rigorously tested during development
- Their models were released to production gradually and KPIs were closely monitored.
- When those were deemed satisfactory, humans derived decisions to aggressively expand their purchasing programme.

Any good Data Science team would do that. It's their job and part of deploying to prod.

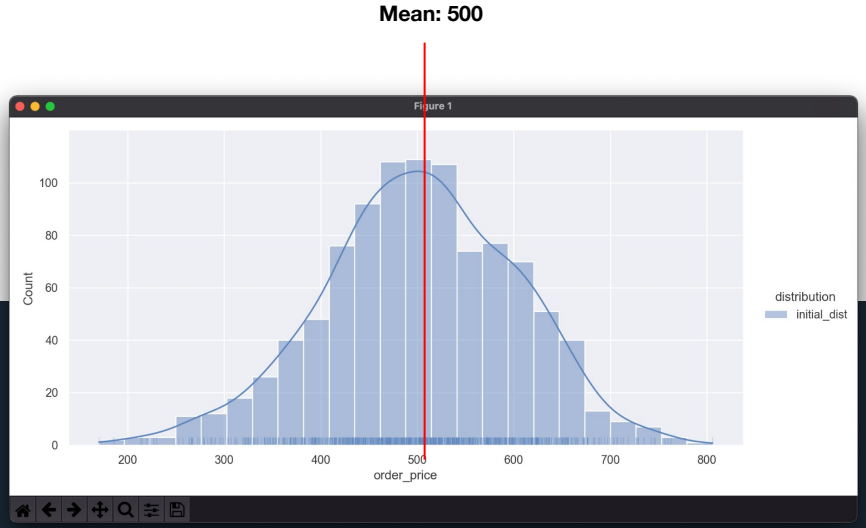
Could it have been avoided?

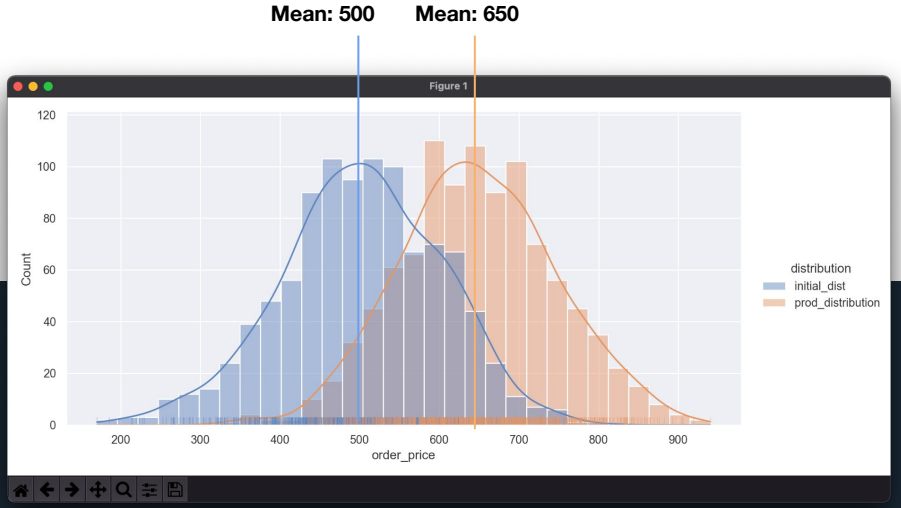
Yes!

- Some phenomena in nature are likely to change, and can do so drastically. When it comes to pricing, that's definitely true.
- This is commonly referred to as "data drift" and it can be detected by:
    - Tracking and alerting on drift.
    - Tracking and altering on accuracy.

Any good Data Science team is aware of that, but data quality management is not their job or core product.

A Detour Into Data Drift





So what is data drift?

When the distribution of one or more of your input features has changed between, for example training time and deployment.

How do you detect it?

01

“Freeze” a reference distribution

02

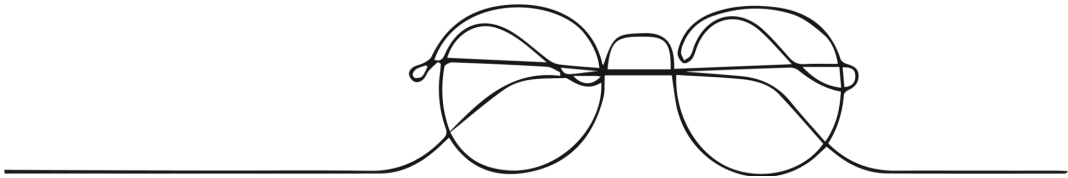
Compare distribution at time t+n and reference distribution

Simple right? 

On the Importance of A Data Quality Framework, Whichever It Is

Why data quality monitoring is “hard”.

- Simply put: you have to write a bunch more code
- Choose your methods from a sometimes large pool
- Orchestrating the checks
- Make it reusable
- Maintain and extend
- The list goes on…



© Soda Data NV. Proprietary and Confidential.

Data quality should not add time to release.

- Developing ML automation takes time and resources
- Data quality monitoring, isn't an internal data team's core product.
- Implementing data quality monitoring can easily increase the scope of any data product's feature set with no direct value add.
- It often ends up "on the backlog"

© Soda Data NV. Proprietary and Confidential.

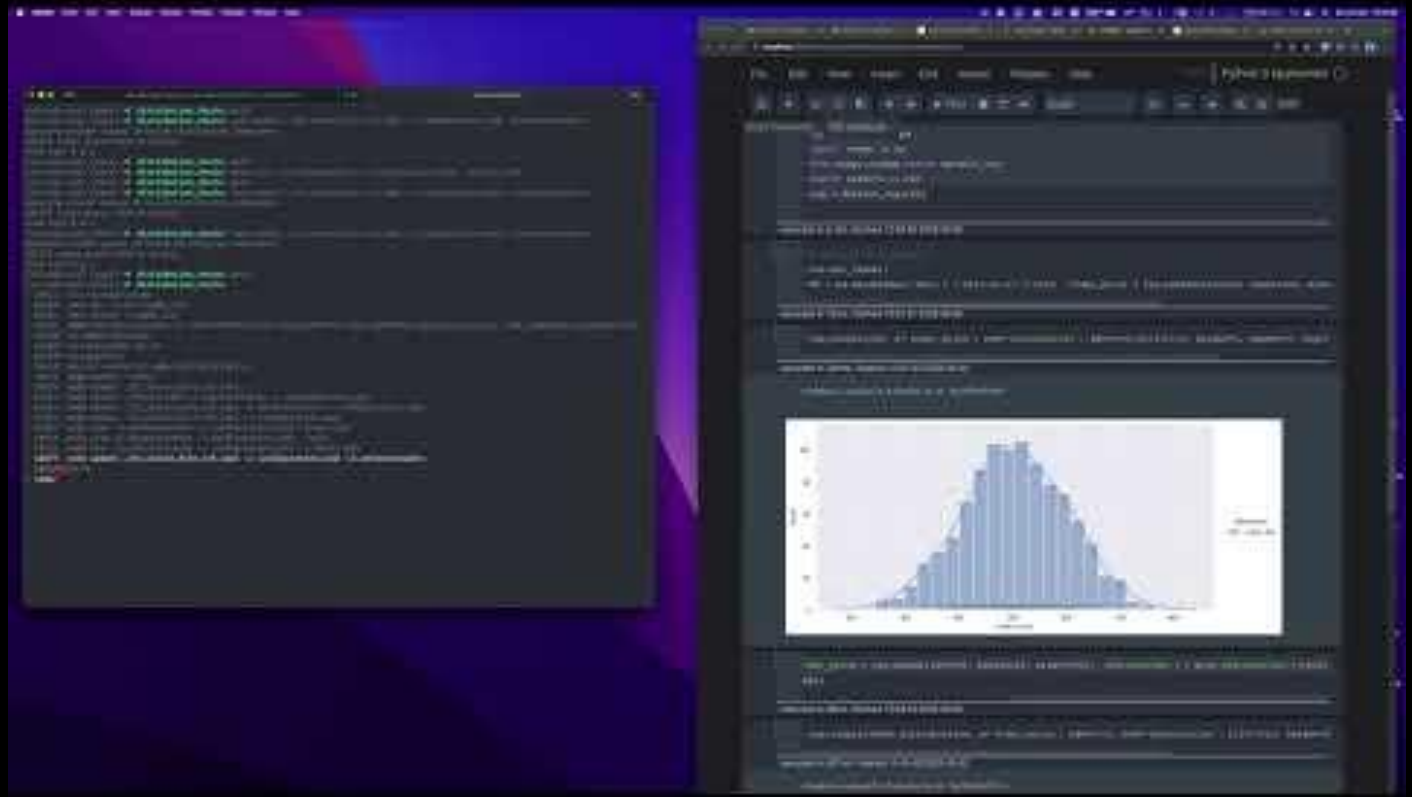
Wouldn’t it Be Nice If…

Wouldn’t it be nice if you could to the following:

```
checks for orders:
  - distribution_difference(order_price, my_happy_ml_model_ditribution) > 0.05:
      method: ks
      distribution reference file: ./orders_order_price_distrib_ref.yml
```


© Soda Data NV. Proprietary and Confidential.

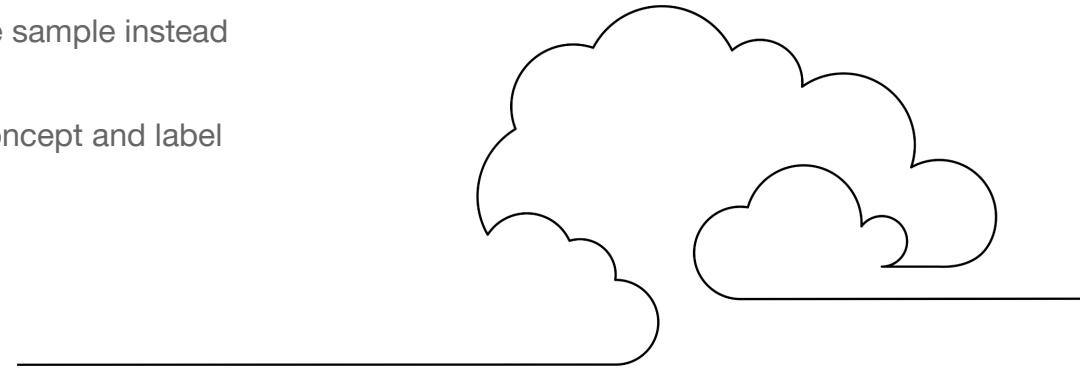
VIDEO



What’s Next?

Why stop there?

- Connect to **Soda Cloud** (to avoid inconvenience of experimental file-based experimental feature)
- Rich visualisation in Cloud/and OSS
- More user control over algos + more algos to choose from
- Entirely data based solution (store reference sample instead of object in cloud/s3)
- Bespoke drift wrappers (monitor for both concept and label drift over one or several datasets)



© Soda Data NV. Proprietary and Confidential.

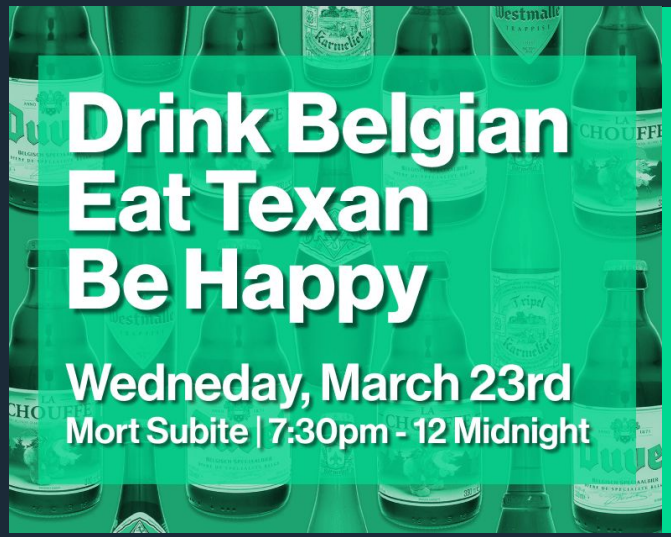
Give it a try!

- docs.soda.io
- `pip install soda-core-[datasource_type] soda-core-scientific`
- https://github.com/sodadata/soda-core

Hit me up, I'll be around!

- Bastien Boutonnet (find me on the socials)
- bastien@soda.io
- www.bastienboutonnet.com

© Soda Data NV. Proprietary and Confidential.



Say Hello While in Austin

- **We're at Booth 23**
- Watch a Soda product demo
- Join our Happy Hour
- Get good swag

Thank You!