Data Contracts in action powered by Python open source ecosystem

Agenda

- About me
- My challenges before data contracts
- Data contracts primer
- Streaming solution with a data contract
- Batch solution with a data contract
- Talk to me

About me

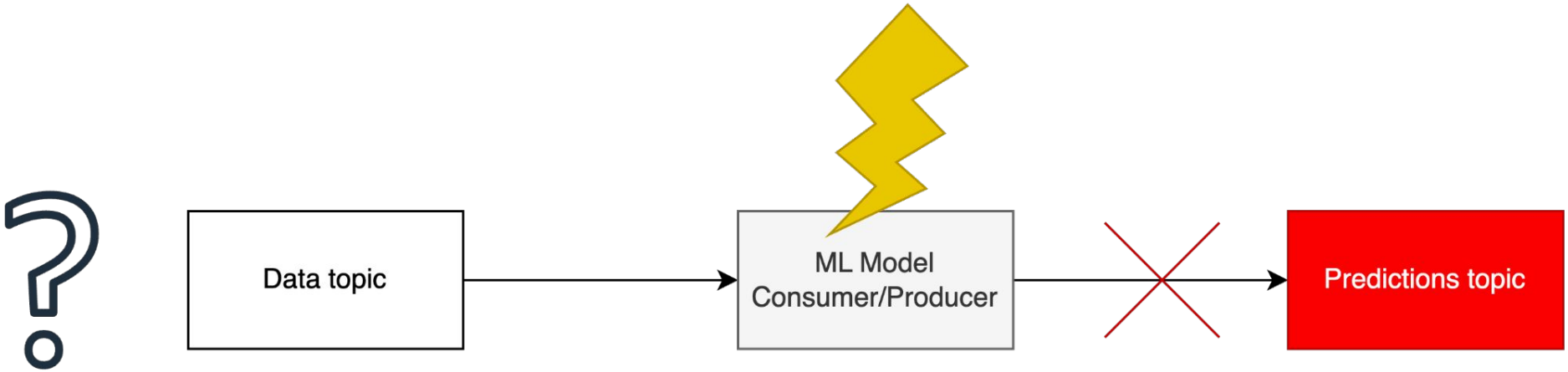
Principal MLOps and Data Engineer at Thoughtworks

Organiser of PyLadies Amsterdam

Microsoft AI MVP since 2021

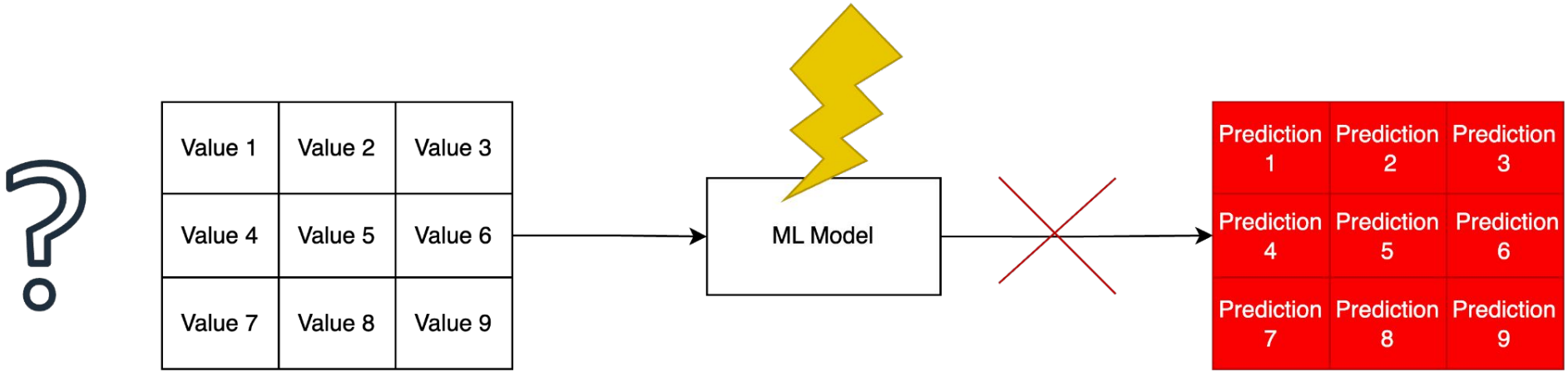
My challenges before data contracts

Streaming challenge



My challenges before data contracts

Batch challenge



Data contracts primer

"A data contract is an agreed interface between the generators of data and its consumers. It sets the expectations around that data, defines how it should be governed, and facilitates the explicit generation of quality data that meets the business requirements."

Andrew Jones, Driving Data Quality with Data Contracts

Data contracts primer

interface

abstraction

agreement

confidence

Andrew Jones, Driving Data Quality with Data Contracts

Data contracts primer

expectations

the structure/schema

data quality checks

SLOs (completeness, timeliness, availability)

ownership and responsibilities

Andrew Jones, Driving Data Quality with Data Contracts

Data contracts primer

governed

PII
sensitivity

access

retention

anonymization

lineage

Andrew Jones, Driving Data Quality with Data Contracts

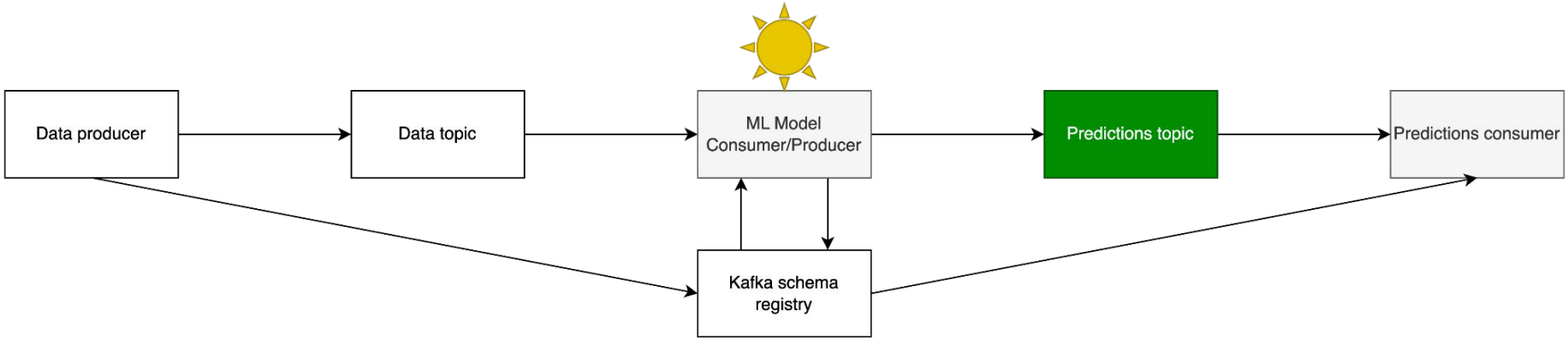
Data contracts primer

explicit

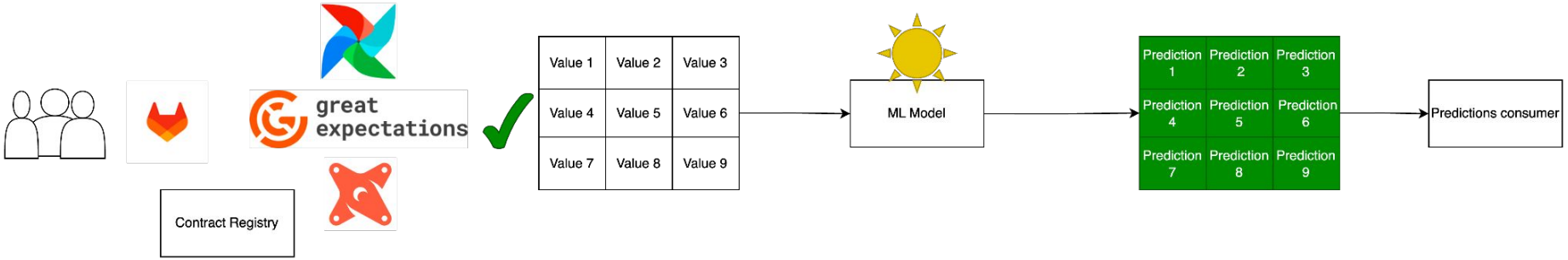
built for consumption

Andrew Jones, Driving Data Quality with Data Contracts

Streaming solution with a data contract



Batch solution with a data contract



Talk to me

About me

About PyLadies Amsterdam