SPARK SUMMIT

PRODUCTIONIZING BEHAVIOURAL FEATURES FOR MACHINE LEARNING WITH SPARK STREAMING

Roman Studenikin & Ben Teeuwen

Booking.com | Priceline Group

#EUstr4

Timeline

2014: Joined Booking.com

Data analysis through Perl

Push for R & python

My first big ML project

Early 2015: Towards production

R randomForest ported

Counter-based features implemented

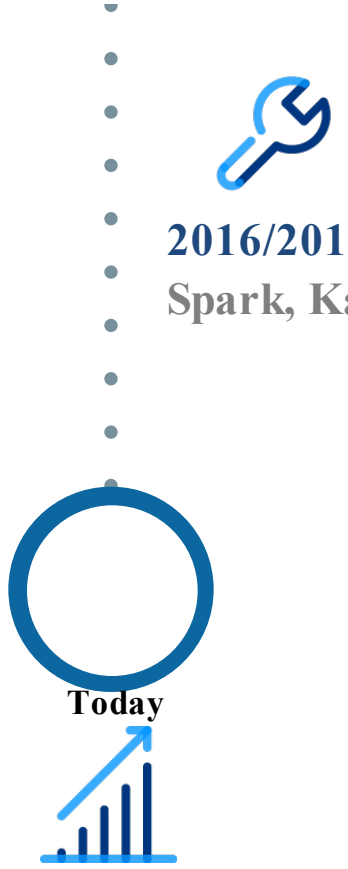
Mid 2015: Validating real-time service

Online predictions

Feature inputs

> Late 2015: Experimentation

.. profit!



2016/2017 continued DS toolkit expansion

Spark, Kafka, Spark Streaming, Cassandra

Today

Today

Integration of components into Feature Store

Trainings, trainings & trainings

Objectives

Reduce skew between offline DS world & online DEV-guarded world

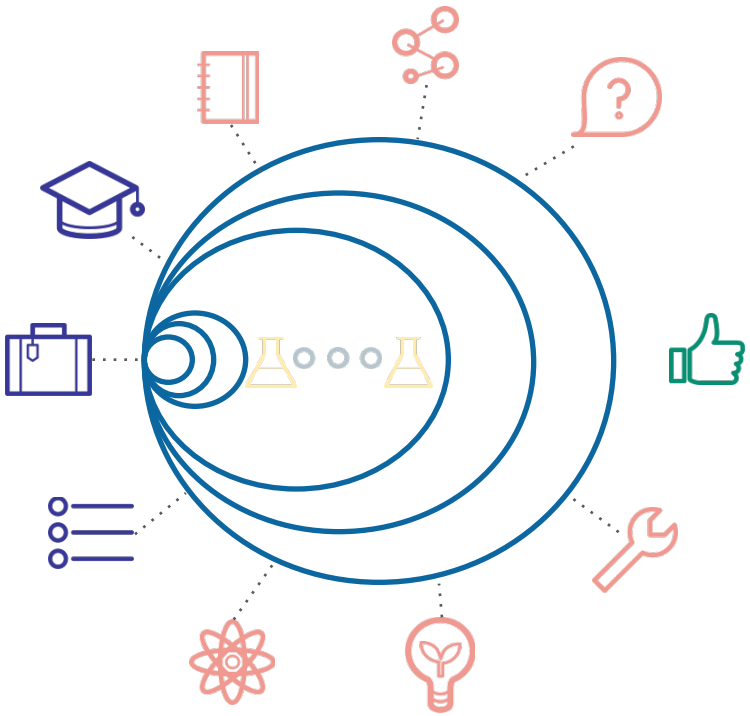
Autonomy for DS (fueled by scarcity of DEV capacity)

Re-usage of DS products

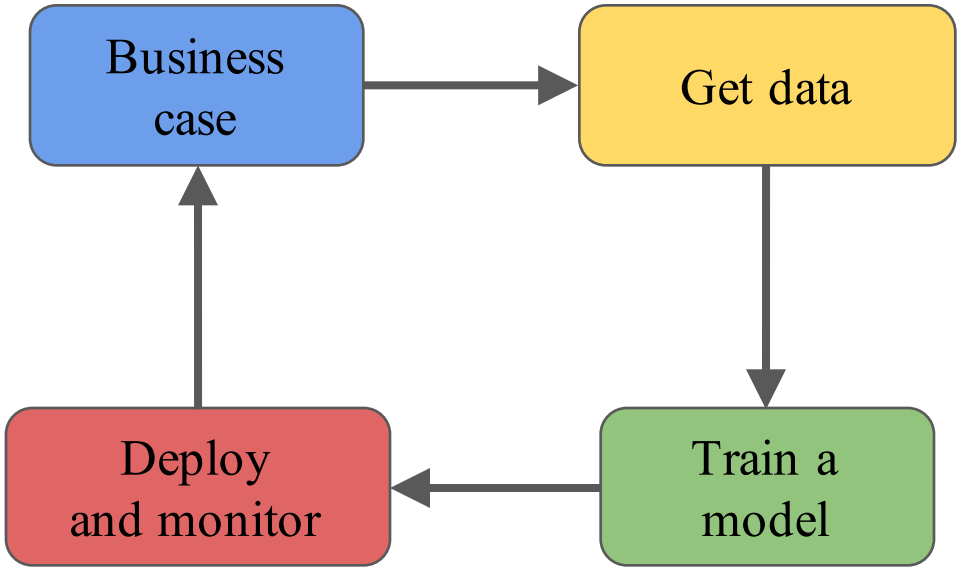
Speed up experimentation cycle with product owners


Our tooling

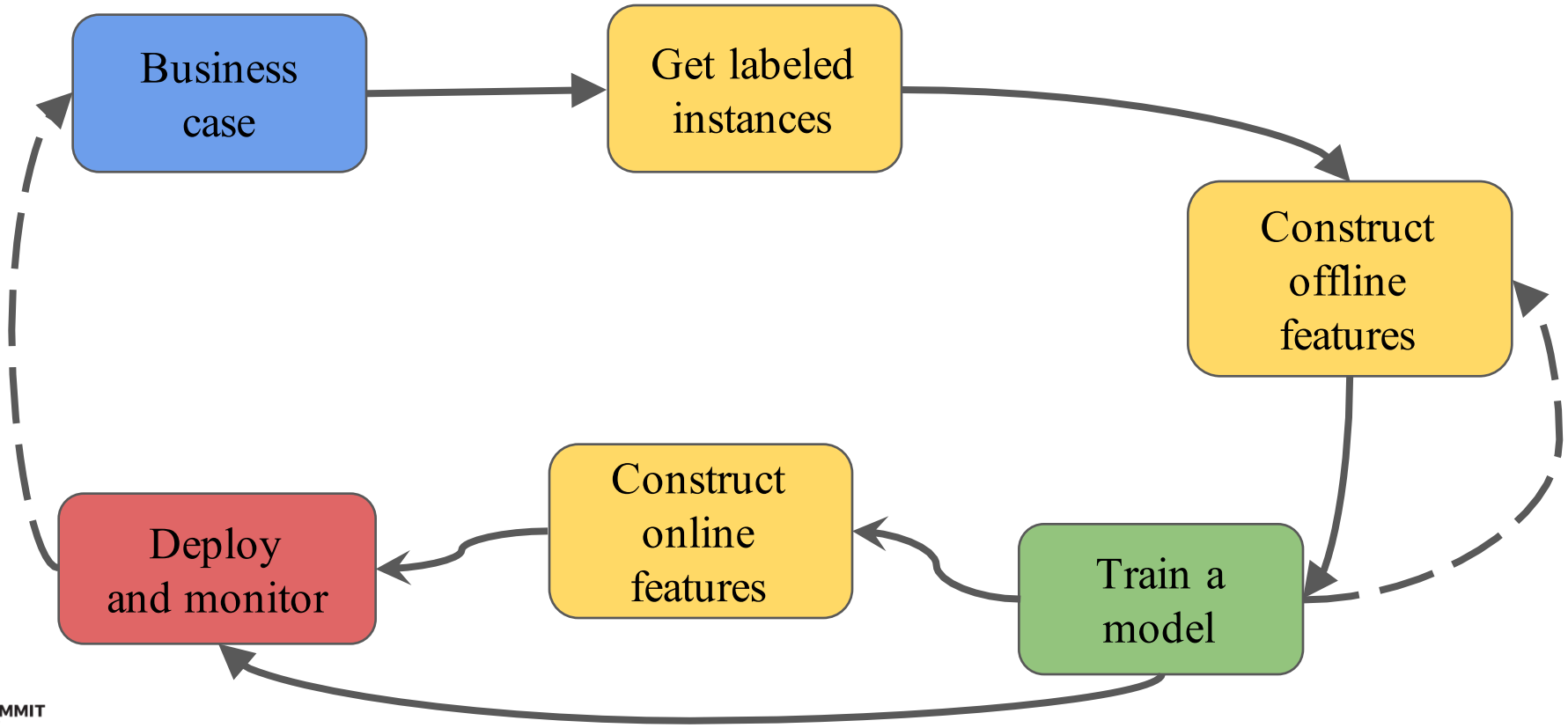
Software development



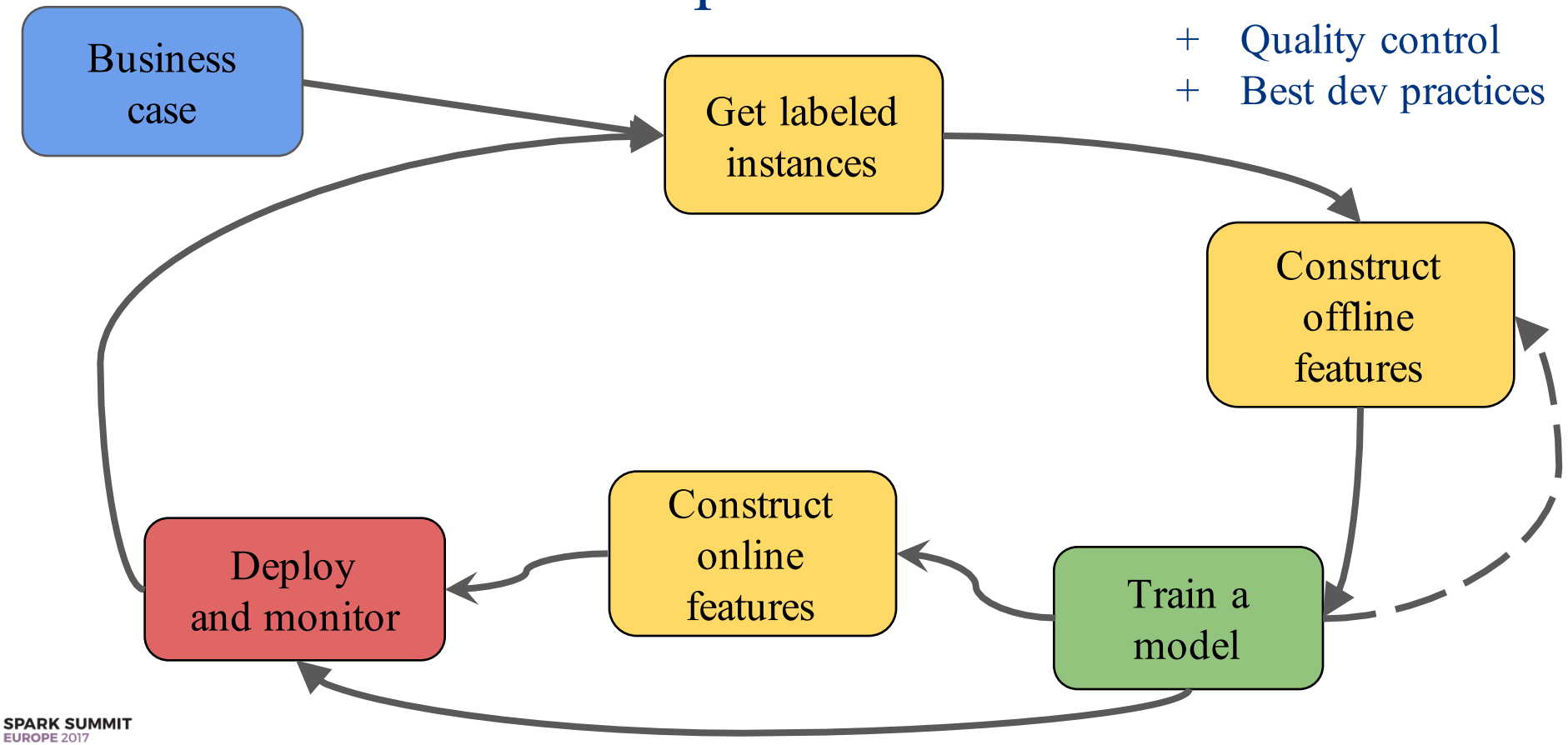
Data science



What can be improved?



Reproducible



Construct offline features

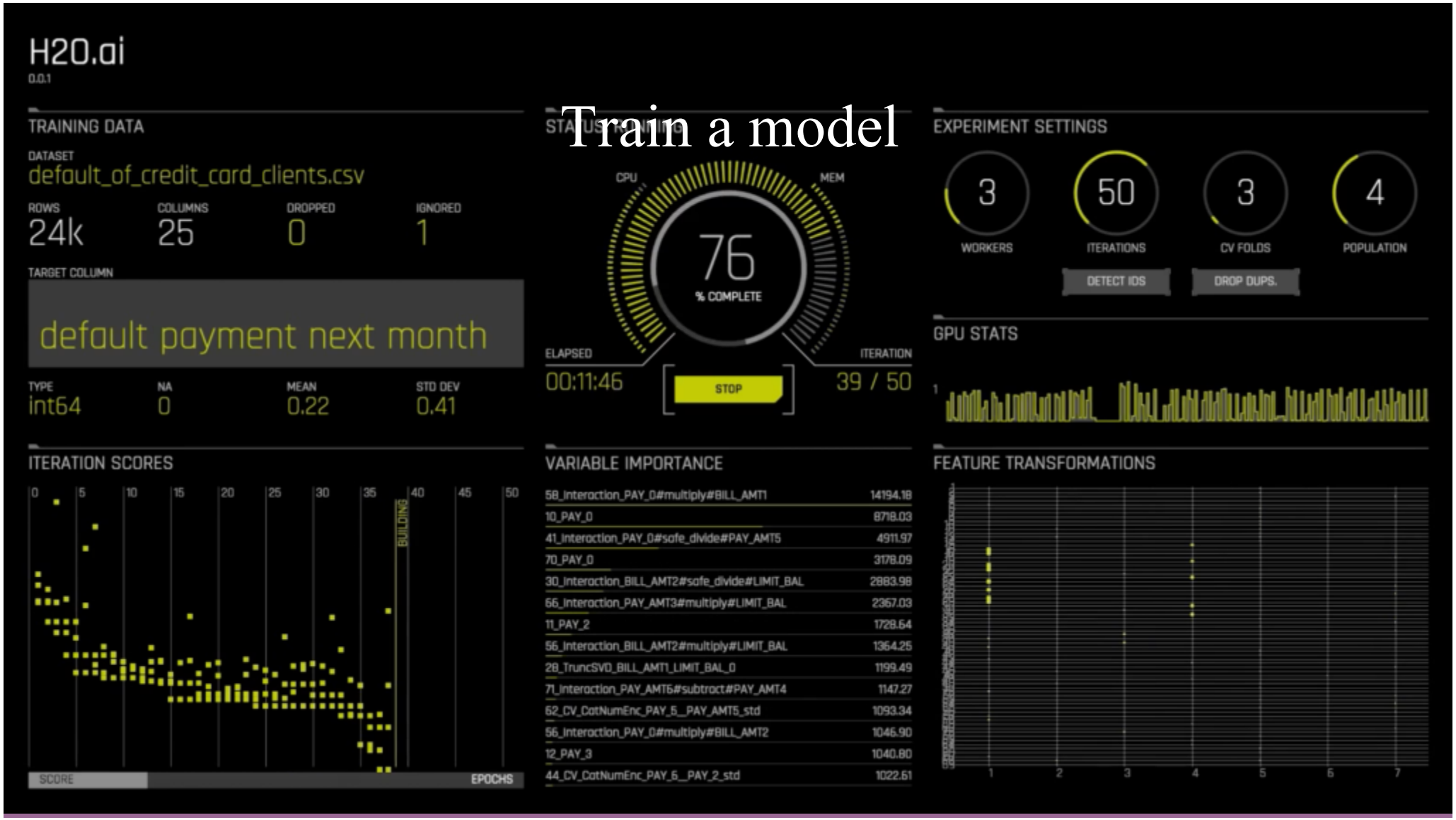
- FeatureVader - registry of all features
- Spark ML Transformer

```
data = spark.sql(....)
fv = FeatureVader()
fv.setNeededColumns([“feature1”, “feature2”])
withFeatures = fv.transform(data)
```

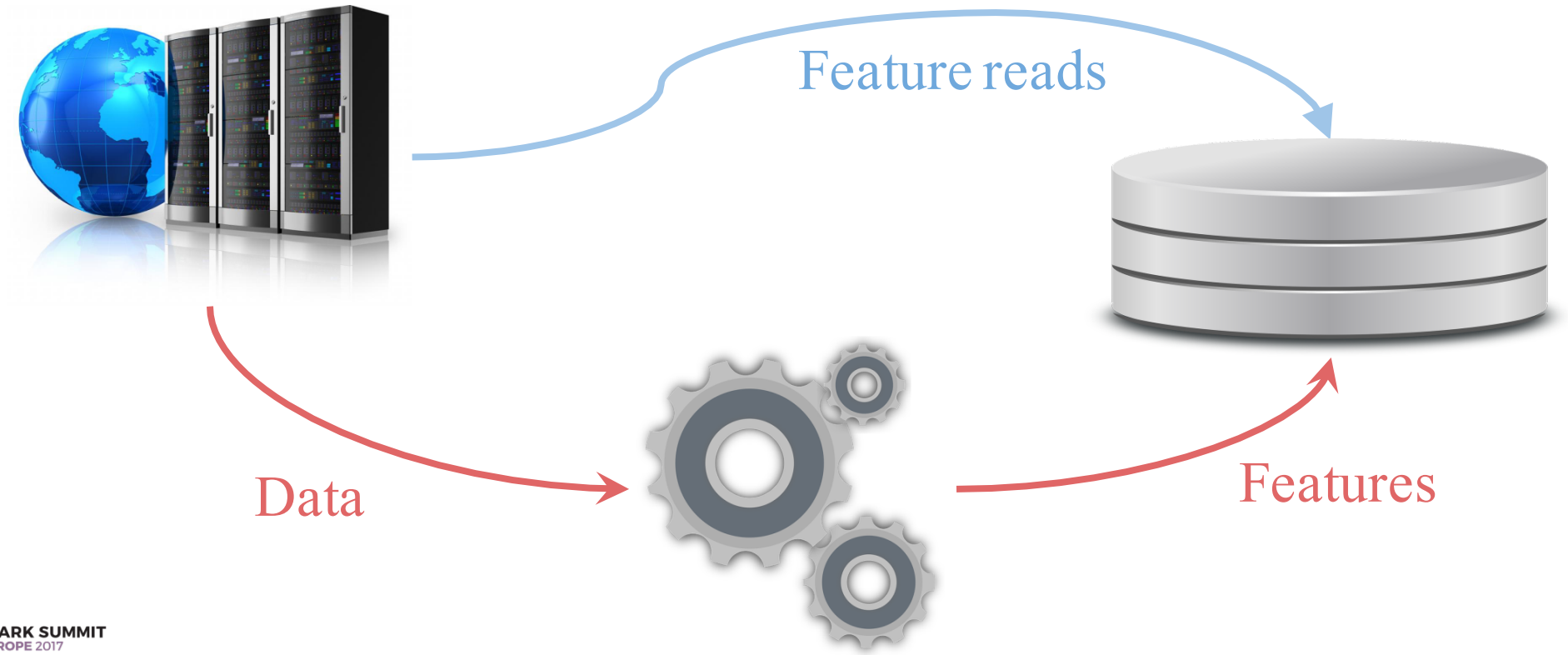
userId	unixTime	label
1001	1508083925	1
1002	1508189936	0

userId	unixTime	label	feature1	feature2
1001	1508083925	1	123.2	2 kids
1002	1508189936	0	213.5	0 kids

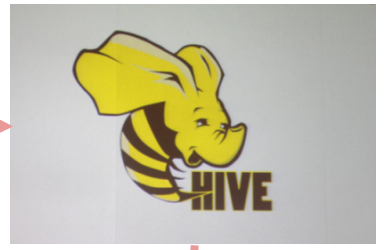
Train a model



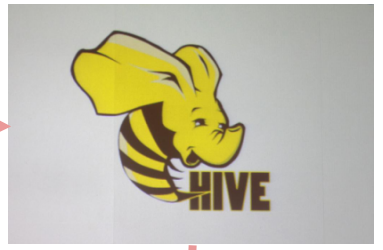
Construct online features



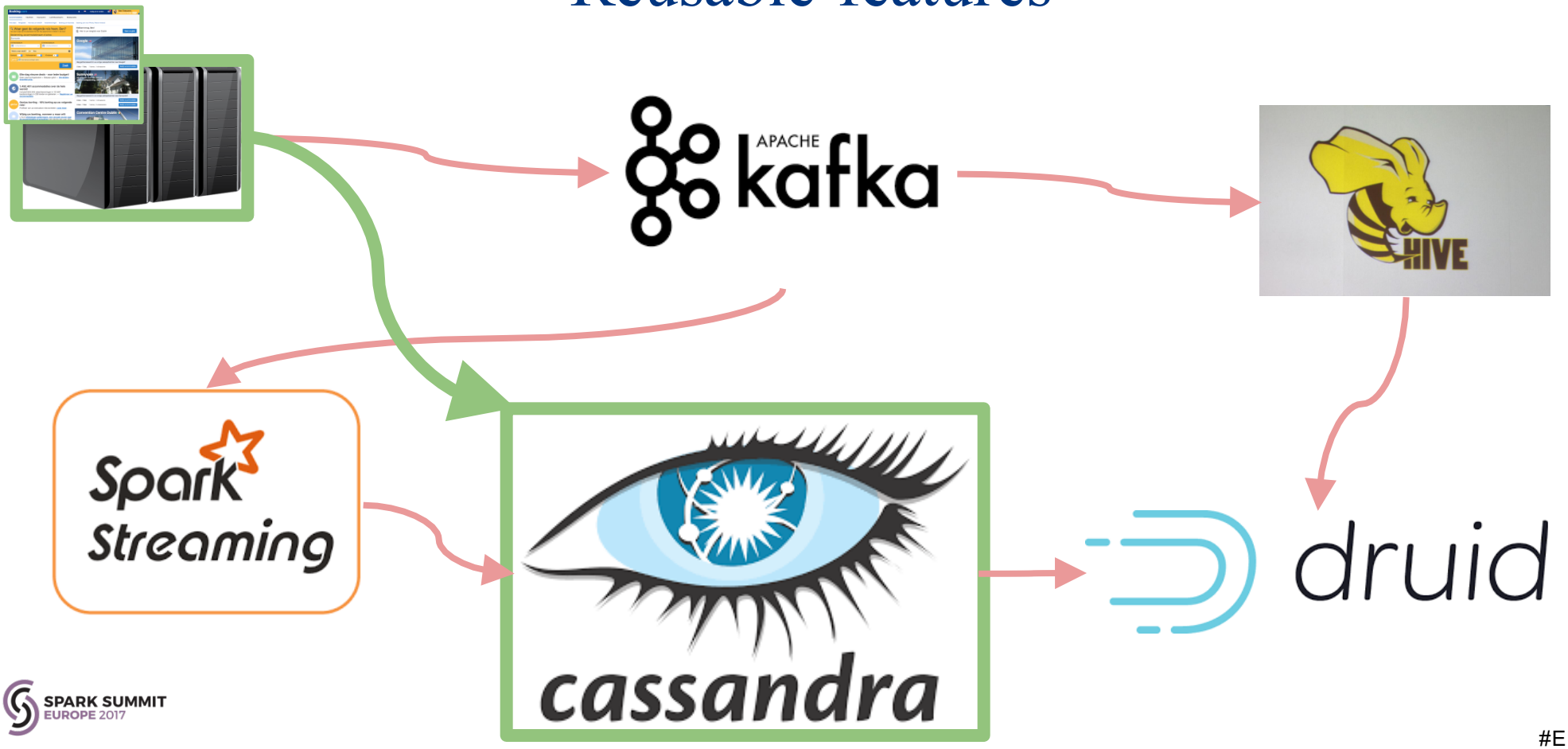
Online architecture



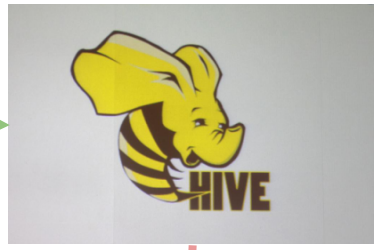
Stable pipeline - at least once semantics



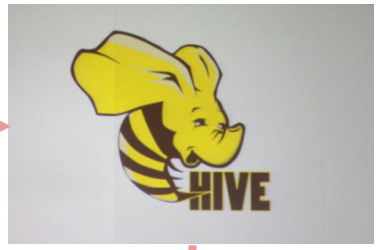
Reusable features



Reduce Training-Serving skew



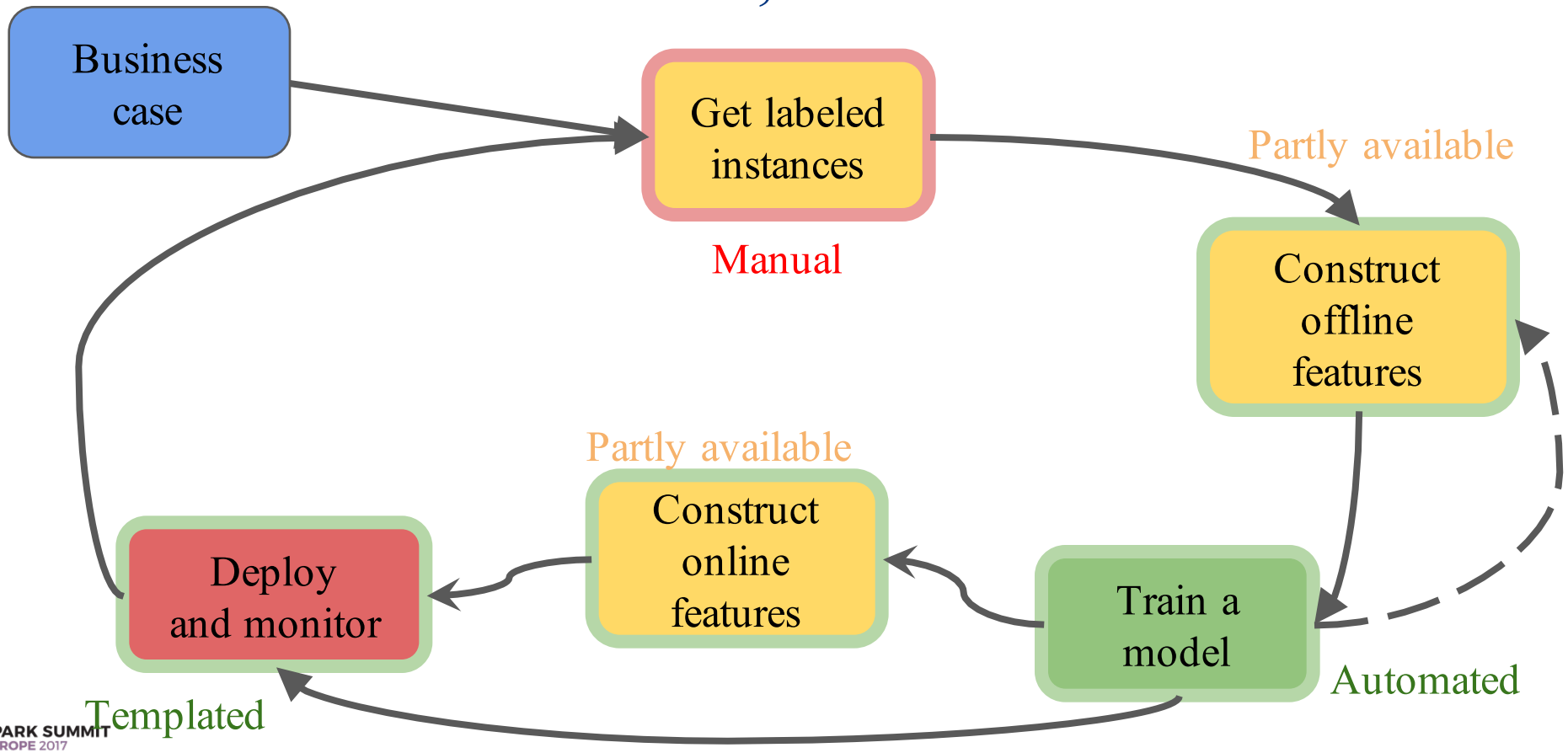
Can be done by Data Scientist



druid

Online feature encoder

Automated, standardised



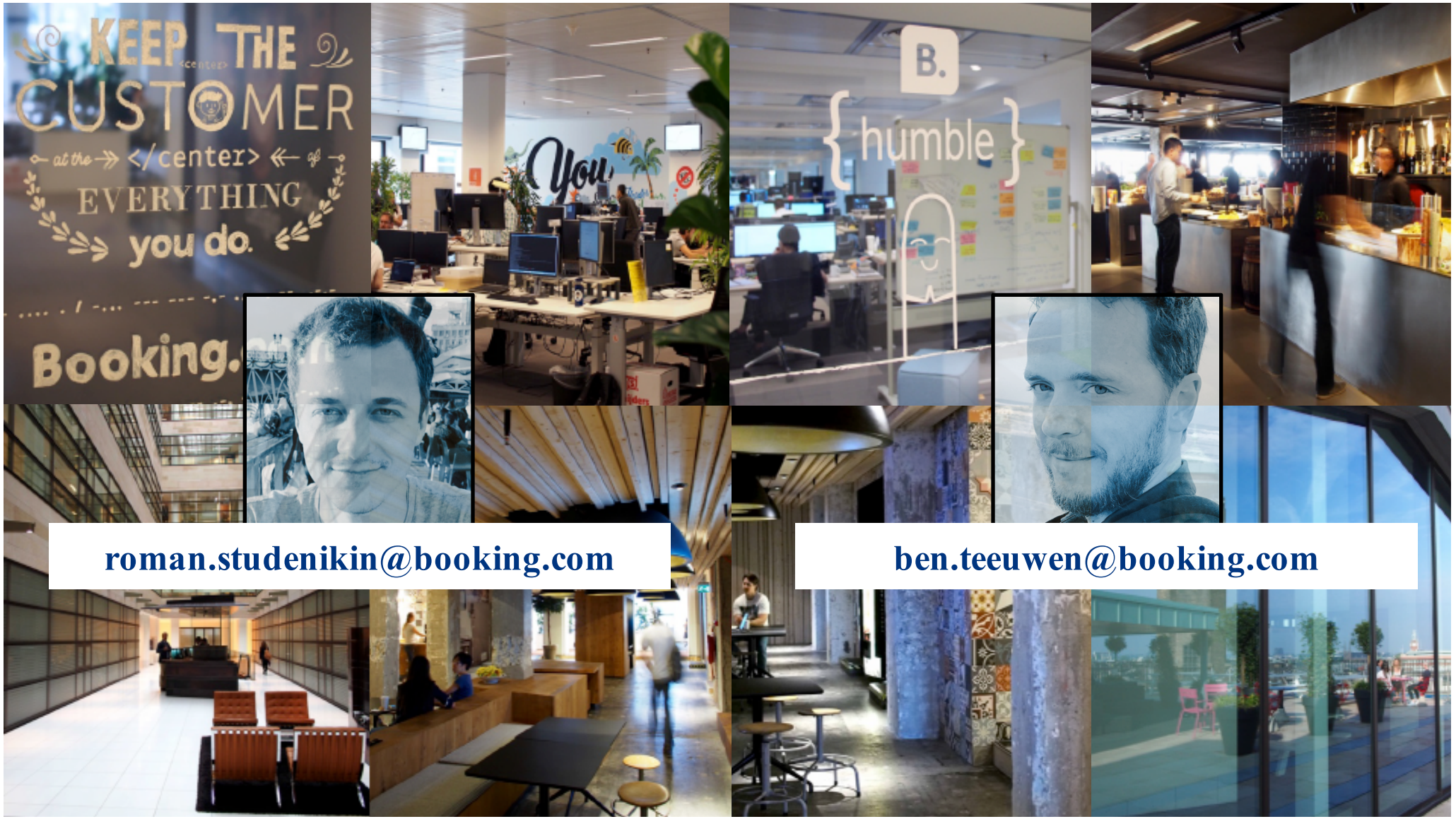
Learnings

Yes, it is possible =)

Offline (simple) feature construction is still easier.

Give data, not tooling.

Start building your vision from something



roman.studenikin@booking.com

ben.teeuwen@booking.com