TECH WEEK

KAIZEN

Andrea Baita

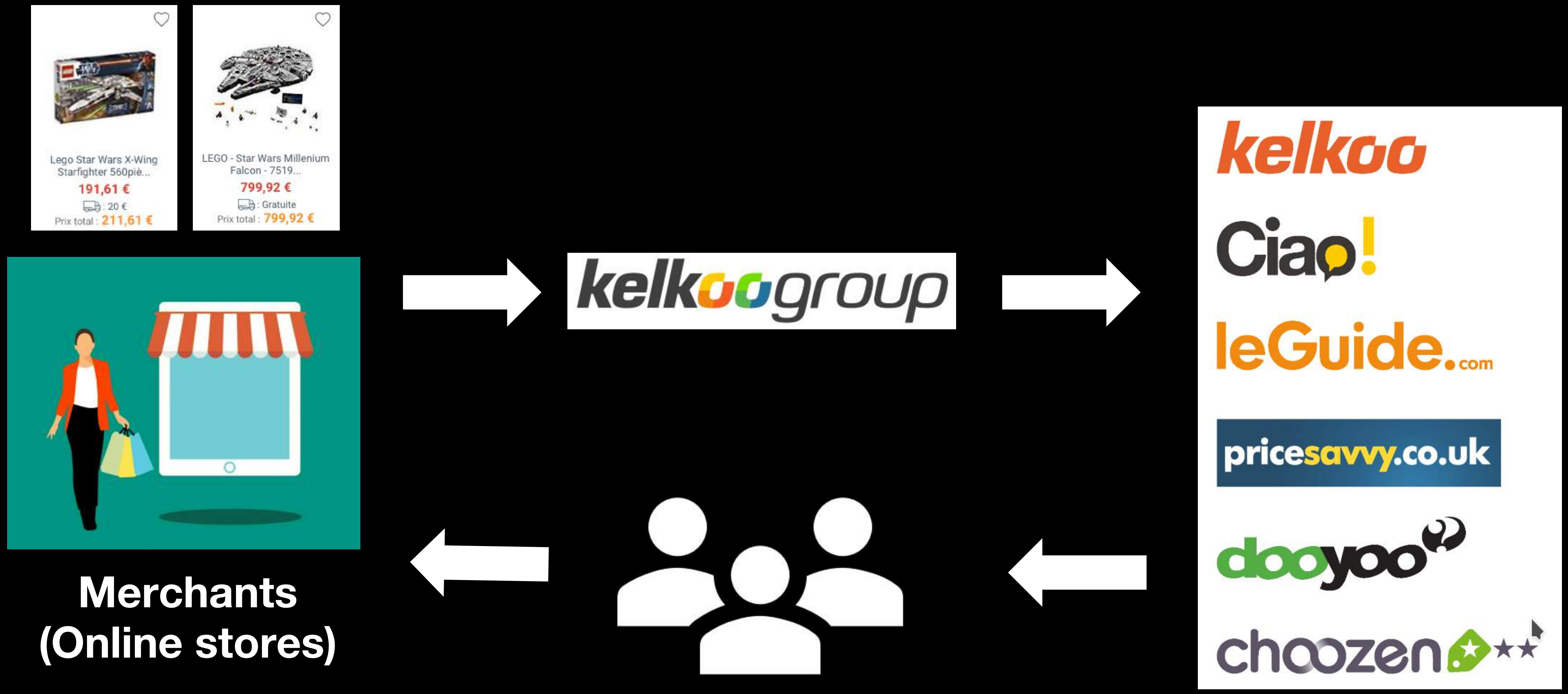
Machine Learning avec Spark : La voie de la production

André Bois-Crettez

Plan

- Business case
- Machine learning
- Testing
- Production
- Lessons learned

Business Case



Merchants
(Online stores)

Users

+
Publishers
(Ads Network)

Targets

KelkooGroup

- Automatization
- Increase margin

Merchant

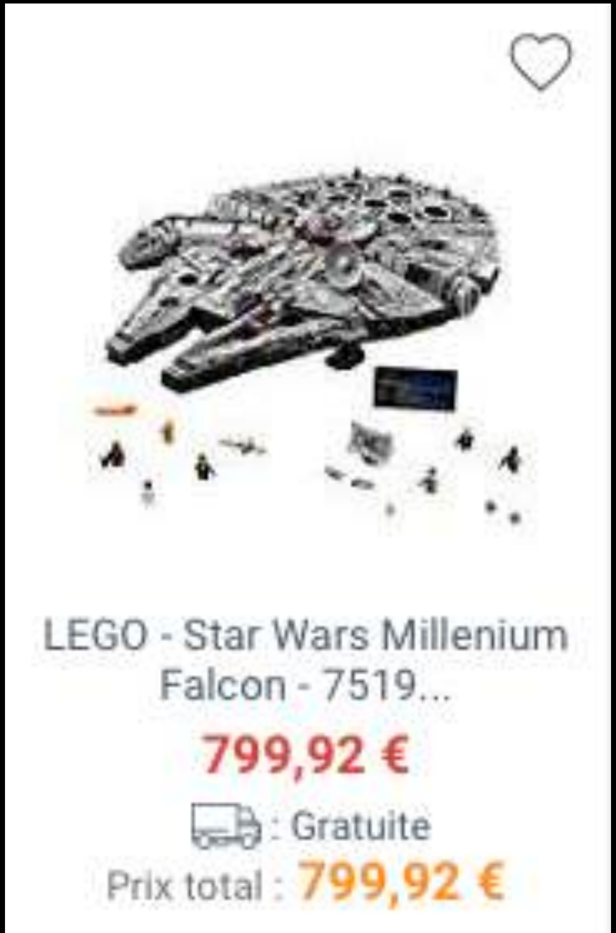
- Attract more buyers
- Sell more with less budget

End Users

- See interesting products
- Find the best offers

Decisions to make

- Where to show the offer (which site, which publisher)
- How much to pay for it



LEGO - Star Wars Millenium Falcon - 7519...

799,92 €

🚚 : Gratuite

Prix total : 799,92 €

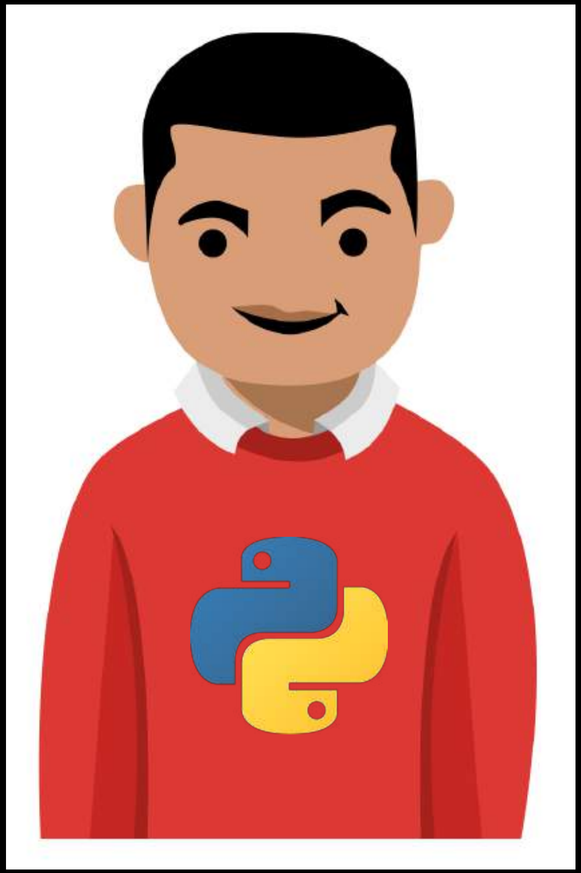
Problem

How many clicks the offer will get ?

Solution

Machine Learning

How?



Data Scientist

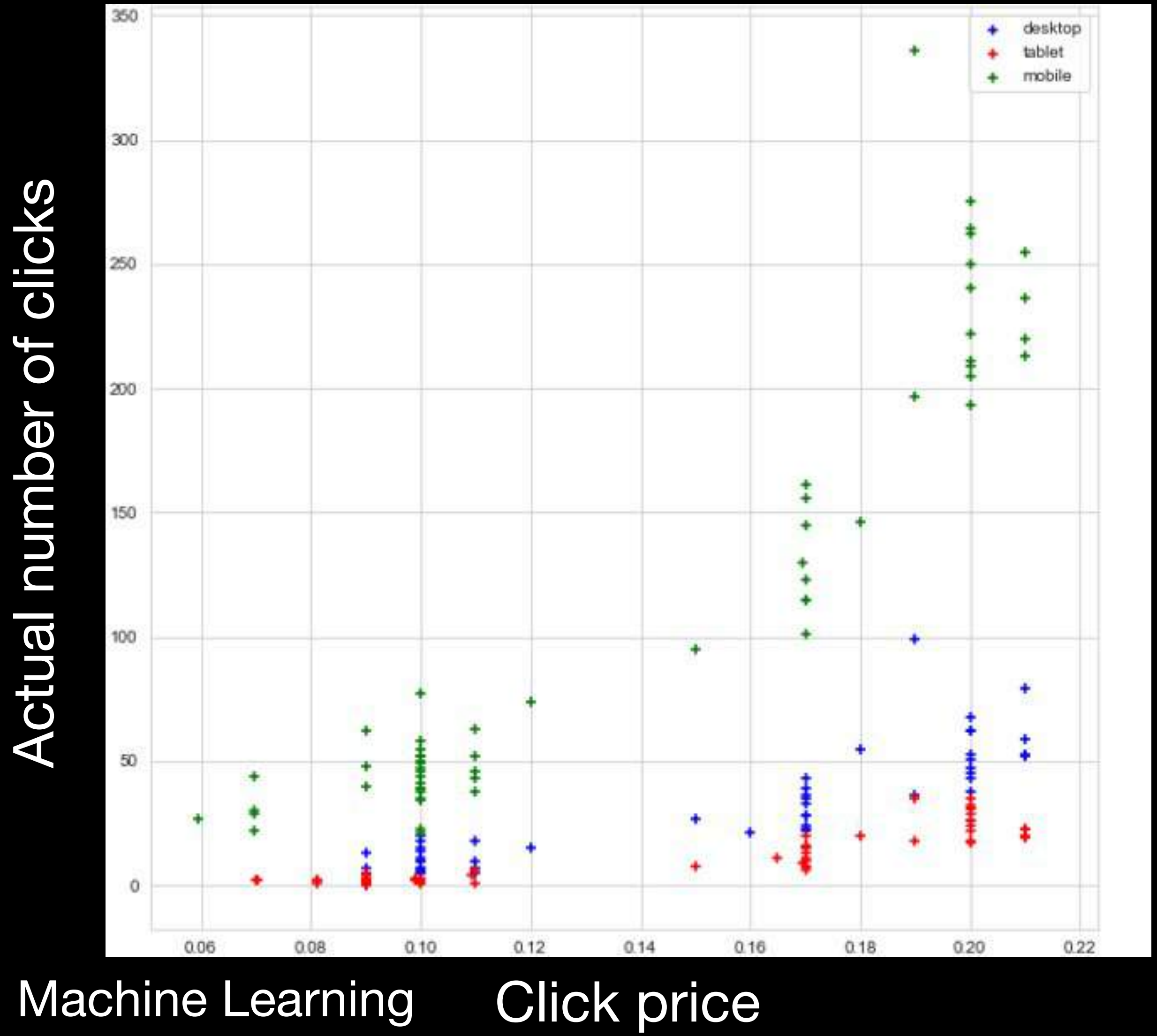
+

Data

=

ML MODEL
(prototype)

Lots of data



Color = type of device

More features are used

time
category
merchant
… secret ones ...

Machine Learning

Learn first ...

Example : past data

date	categoryId	merchantId	category	device	price
08/04/2019	10163969	1	Accessoires Moto	desktop	0.08
08/04/2019	10163969	1	Accessoires Moto	mobile	0.0704
08/04/2019	10163969	1	Accessoires Moto	tablet	0.18
08/04/2019	10543669	2	Lingerie Femme	desktop	0.23
08/04/2019	10543669	2	Lingerie Femme	mobile	0.0989
08/04/2019	12676471	3	Lunettes de vue	mobile	0.1204

with past result

clicks
2
21
22
10
2
1

= Model

… then predict !

Current data

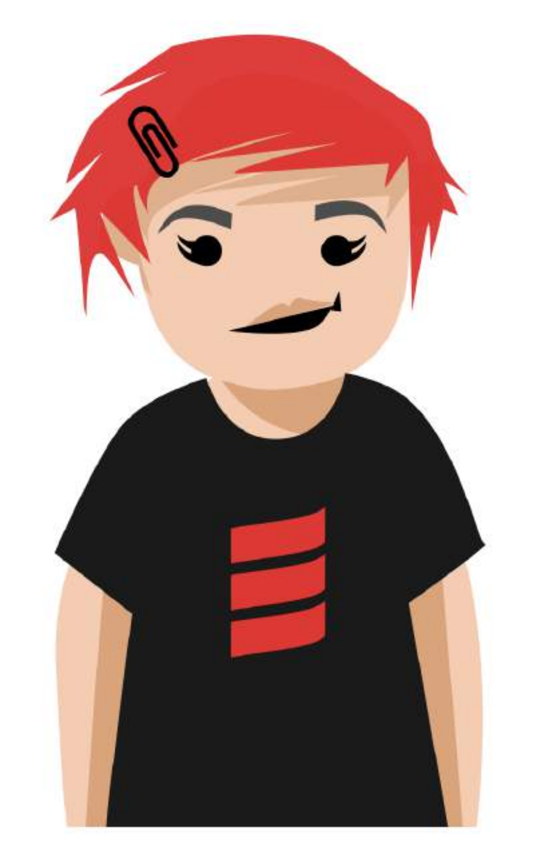
date	categoryId	merchantId	category	device	price
11/04/2019	10163969	1	Accessoires Moto	desktop	0.09
11/04/2019	10163969	1	Accessoires Moto	mobile	0.08
11/04/2019	10163969	1	Accessoires Moto	tablet	0.19
11/04/2019	10543669	2	Lingerie Femme	desktop	0.24
11/04/2019	10543669	2	Lingerie Femme	mobile	0.10
11/04/2019	12676471	3	Lunettes de vue	mobile	0.13

with Model

= Predict result

Predicted clicks
3
20
23
11
1
2

How do we implement it?



Scala Developer

+ APACHE Spark =

ML MODEL
(production ready)

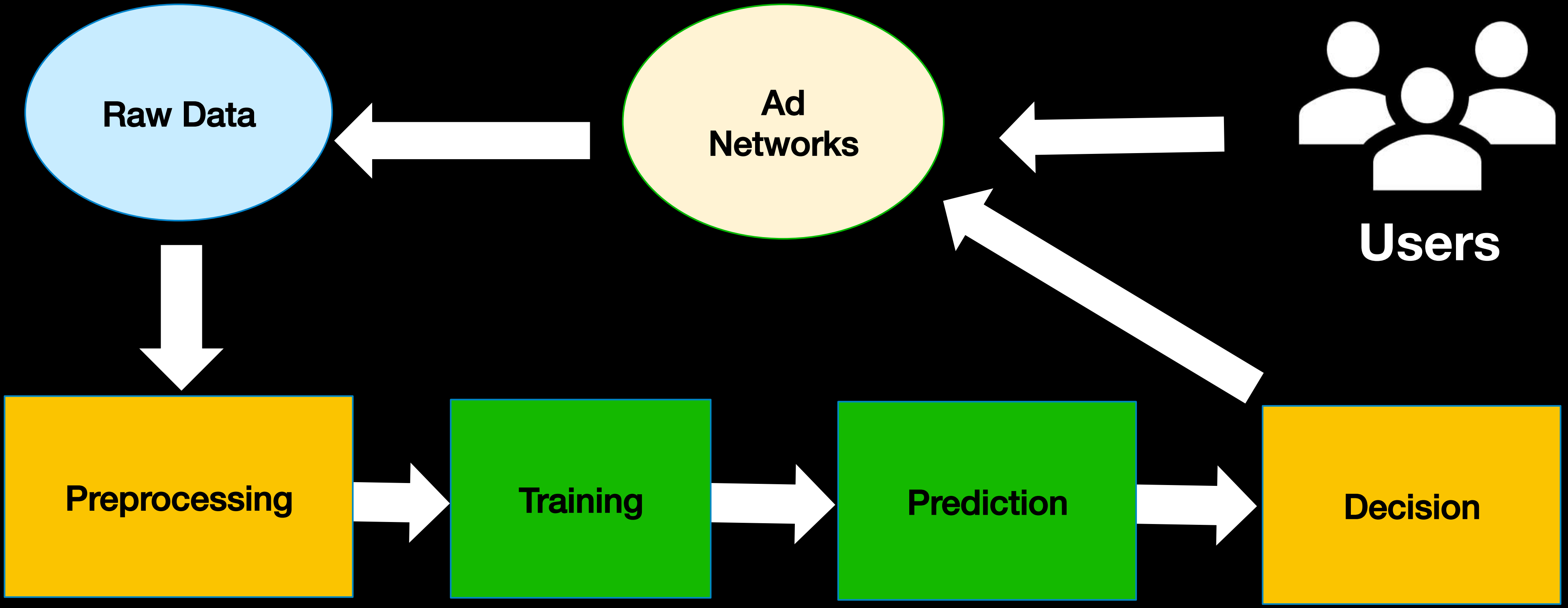
Spark?

Unified analytics engine for large-scale data processing

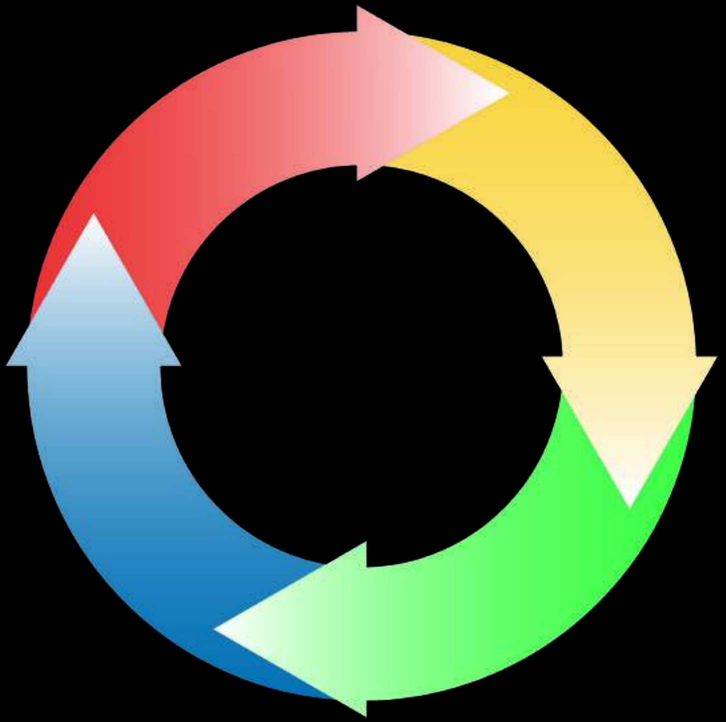
- interactive exploration
- batch processing
- SQL
- machine learning at scale
- ...

How do we use it?

Architecture



Data



Learn

Model

Predict and Decide

The model changes over time ...

… how can we deploy it?

Model deployment approaches

✗

Train first and then deploy the model

- Real time predictions
- Models training is expensive
- Training data is stable

✓

Deploy the code, train at needs

- Batch predictions
- Quick model training
- Training data evolve fast

How can we test it?

ML testing problems

- Behavior depends on data
- Difficult to define exact test result
- Code is hard to structure
- Unit tests are challenging

Solutions

- Compare metrics, not values
- Use functional testing
- Live monitoring
- Tracking over time

How to define the metrics?

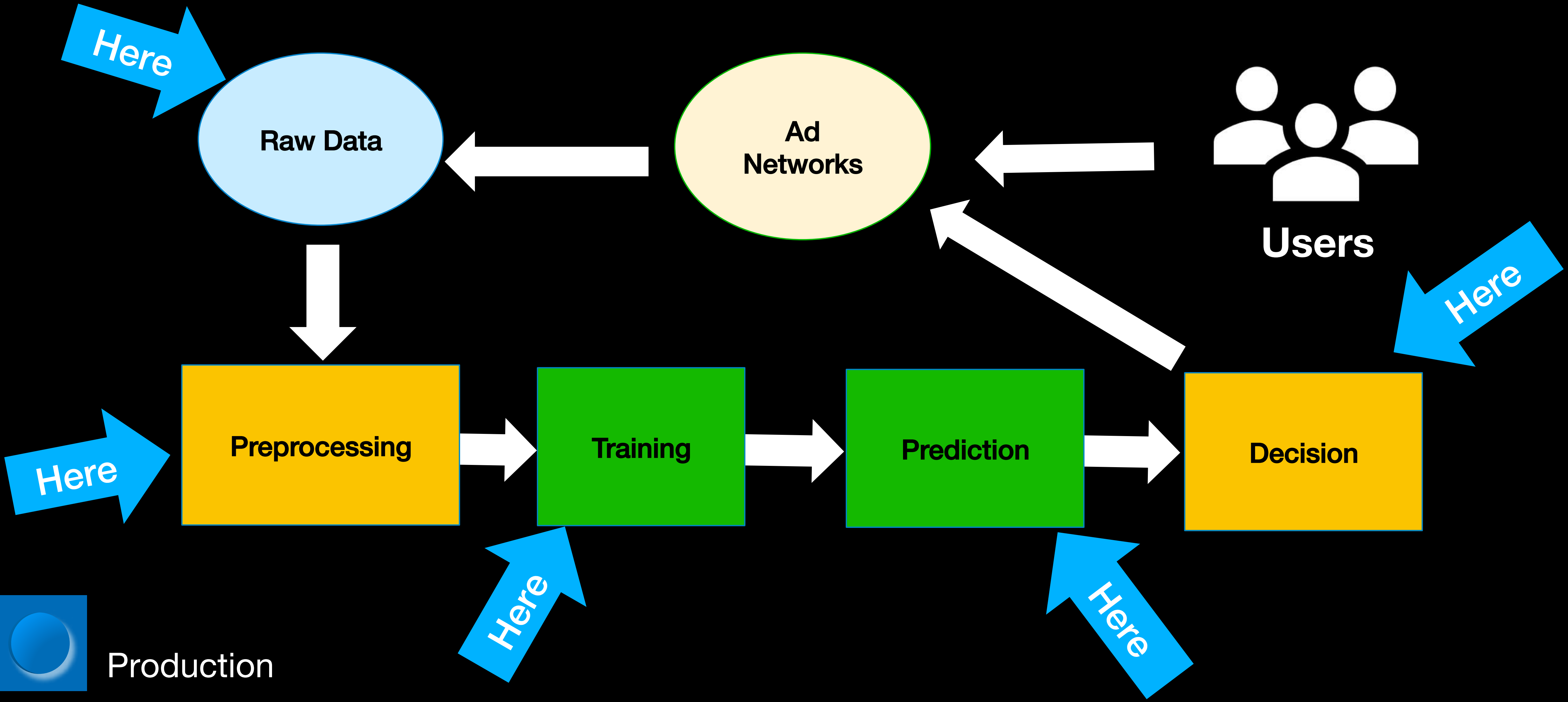
Production

Define relevant metrics

Goal : evaluate quality

- **Prototyping: Statistical metrics**
    - Mean Average Error, Root Mean Square Error
- **Testing: Business metrics**
    - Total margin
- **Monitor: Real time metrics**
    - Predicted Clicks vs. Real Clicks

Tests and Measures : where ?



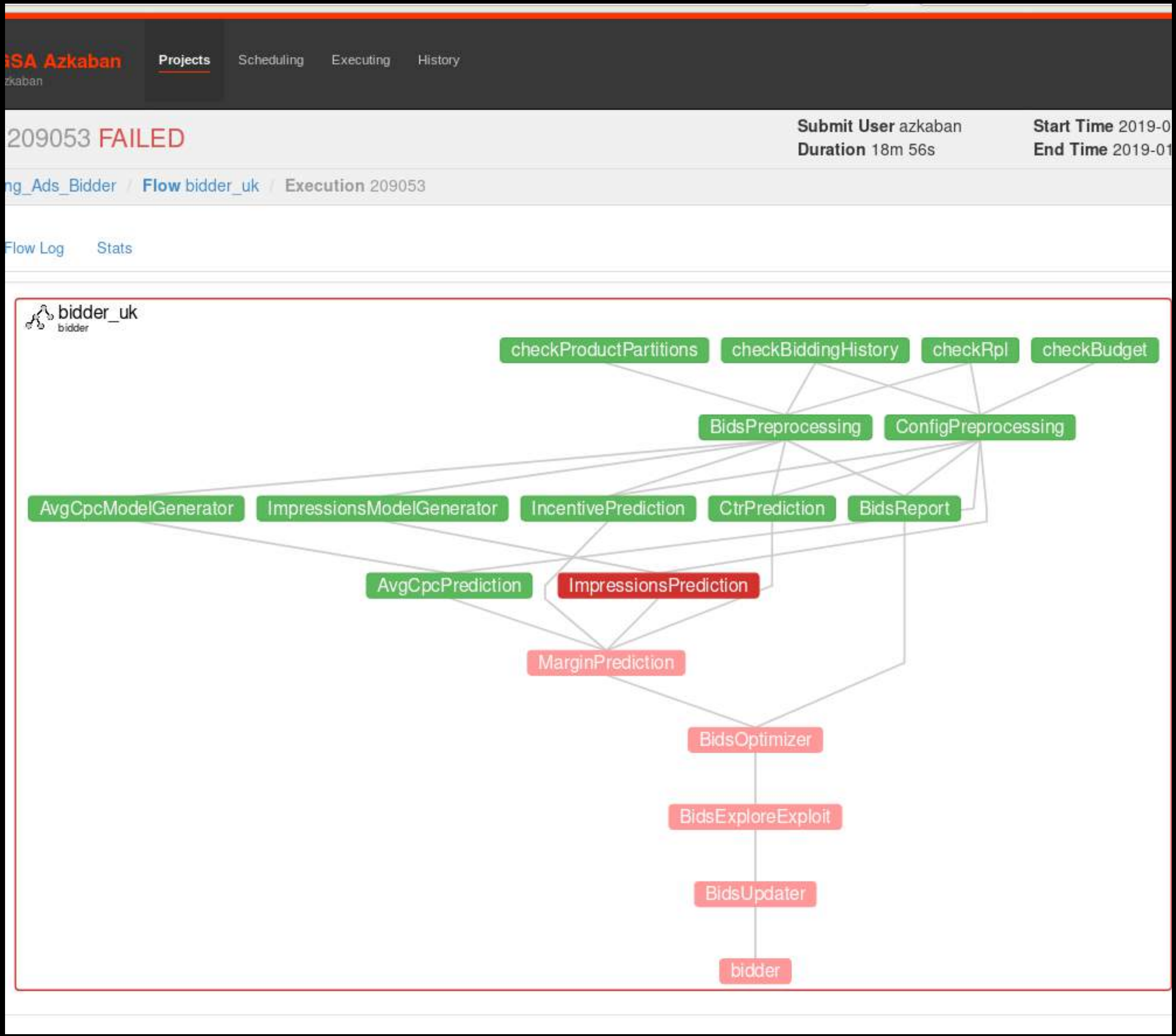
Production

How can we schedule the jobs?

Production

Azkaban

- Workflow job scheduler
- Hadoop and Spark jobs
- Graph of job dependencies
- Alerting on failures with Nagios



209053 FAILED

Submit User azkaban
Duration 18m 56s

Start Time 2019-0
End Time 2019-01

ng_Ads_Bidder / **Flow** bidder_uk / **Execution** 209053

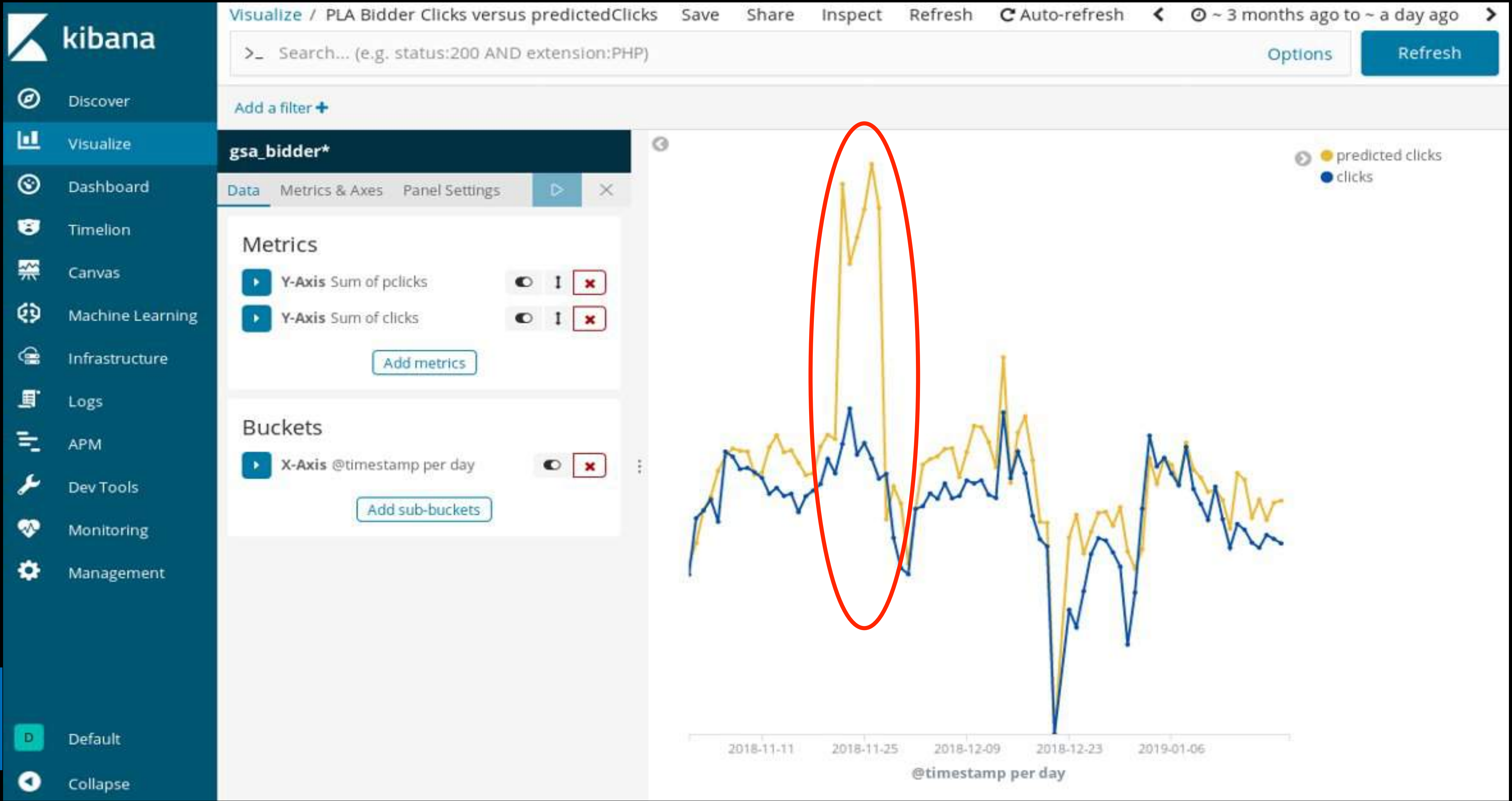
Production

How to track the model behavior?

Tracking

- Business metrics graphs
- Predictions vs. actual results
- Study trends long term
- Adapt model when market changes
  - Easy to fix: abrupt drop in quality metric
  - Harder: slow erosion of quality

Tracking with ELK



Metrics

- Y-Axis Sum of pclicks
- Y-Axis Sum of clicks

Buckets

- X-Axis @timestamp per day

So, what did we learn ?

1

2

I TOLD YOU TO TEST THE MODEL
38
BUT YOU DIDN'T LISTEN
imgflip.com

3

GET FEEDBACK FROM PRODUCTION
TO IMPROVE NEXT VERSIONS
imgflip.com

4

Questions ?