Pragmatic Deep Learning for image labelling - An application to a travel recommendation engine

主讲人：Dataiku数据科学家 Alexandre Hubert

Outline

Introduction and Context

Labeling Images

Pragmatic deep learning for dummies

Iterative building of a recommender system

Post Processing

AKA: Image for BI on steroids

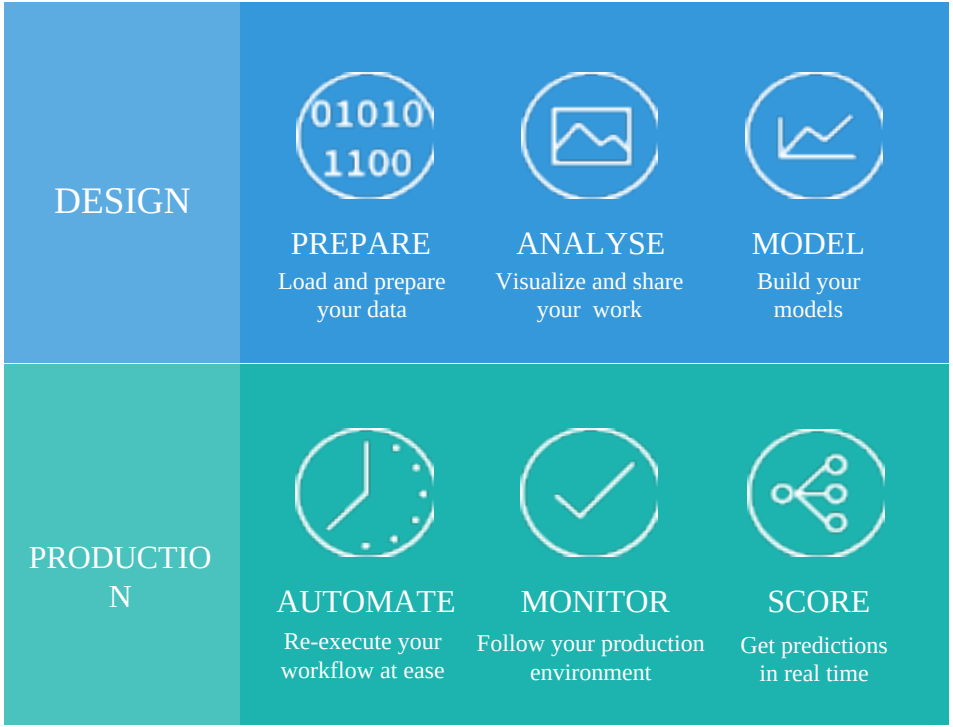
Results

More images !

Dataiku

Data Science Software Editor of *Dataiku DSS*

- Founded in 2013
- 90 + employees, 100 + clients
- Paris, New-York, London, San Francisco, Singapore



data iku

• Key Figures

E-business vacation retailer

Negotiate the best prize for their clients

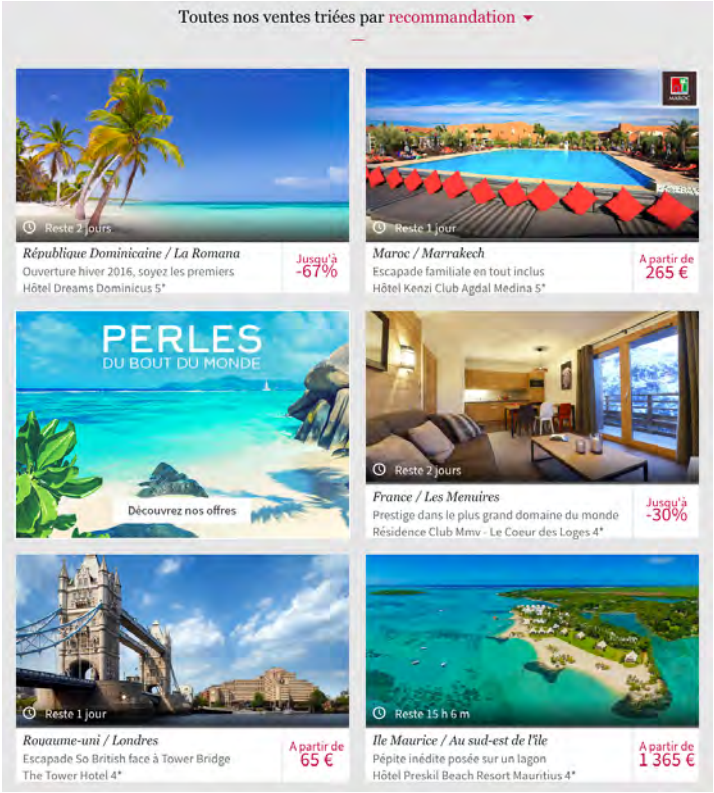
Discount luxury

18 Millions of clients.

Hundreds of sales opened everyday

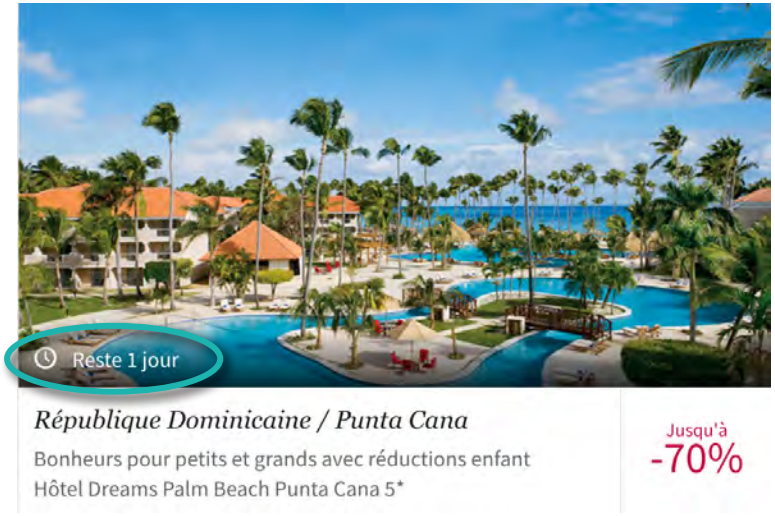
Sale Image is paramount

Purchase is impulsive



data iku

Specificities



Highly temporary sales

-> Classical recommender system fail

-> Time event linked (Christmas, ski, summer)

Expensive Product

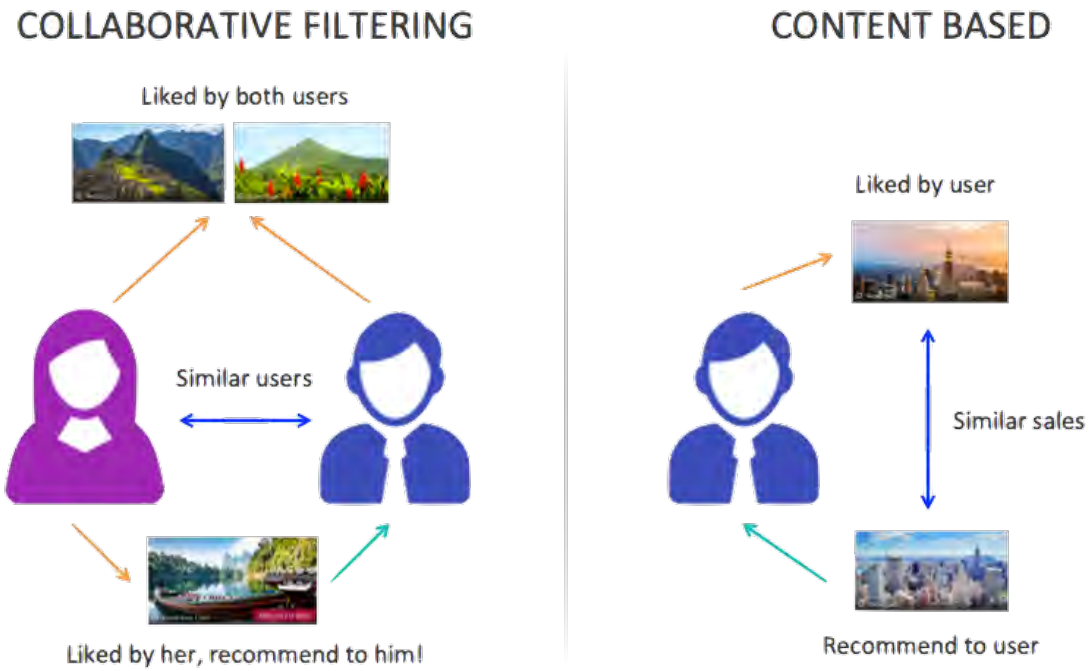
-> Few recurrent buyers

-> Appearance counts a lot

data iku

Iterative Building of a Recommender System

Basic Recommendation Engines



Other Factors

POPULARITY

Liked by many, recommend to all!

RETARGETING

EXTERNAL FACTORS

One Meta Model to Rule Them All

Describe

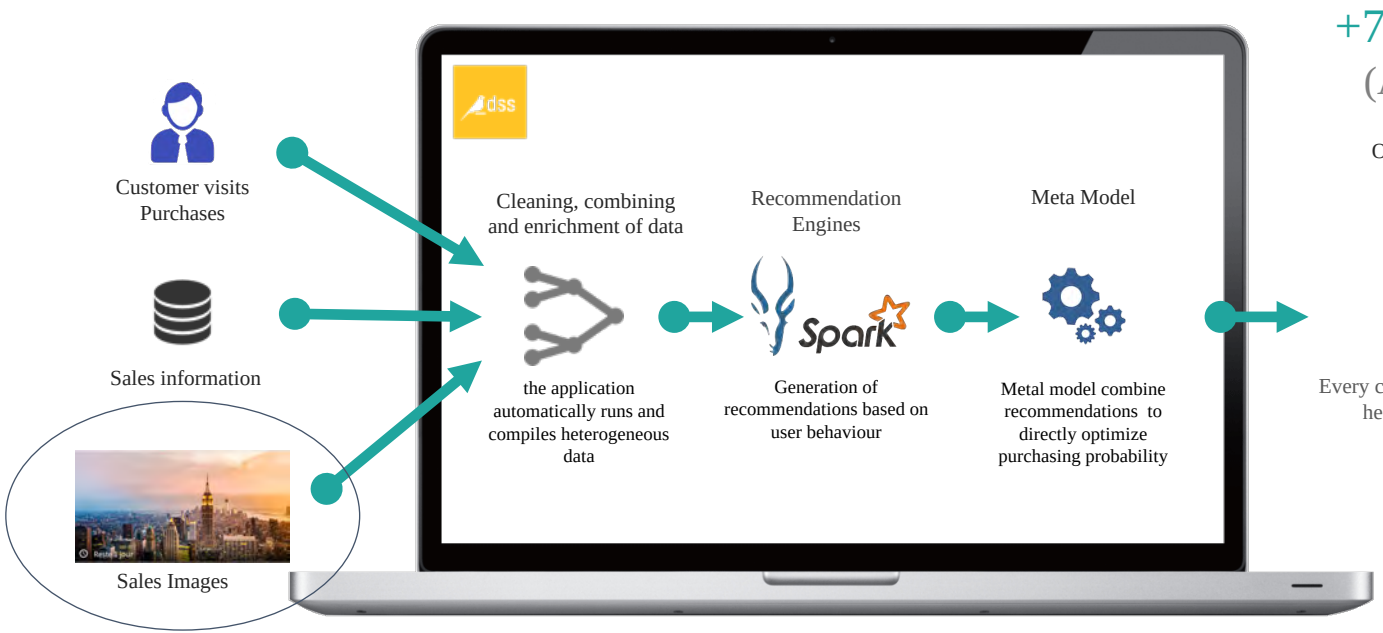
Recommenders as features

Machine learning to optimize purchasing probability

Combine

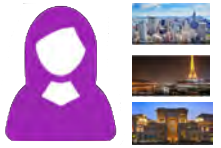
Recommend

Recommender system for Home Page Ordering



+7% revenue
(A/B testing)

Optimization of home
display



Every customer is shown the 10 sales
he is the most likely to buy

↻ Batch Scoring every night

data
iku

Why use Image ?

A picture is worth a thousand words

We want do distinguish

« Ski »

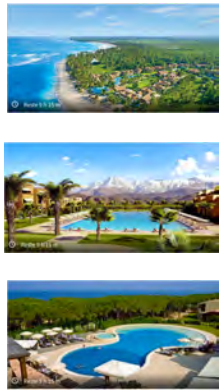


« Sun and Beach »



data iku

Integrating Image Information



Sales Images

Labeling Model

Sea + Beach +Forest + Hotel

Pool + Palm Trees Hotel
+ Mountains

Pool + Forest + Hotel + Sea

Sales descriptions
vector

CONTENT BASED

Recommender
System

Image Labelling For Recommendation Engine

Pragmatic Deep learning for “Dummies”

data iku

Using Deep Learning models

Common Issues

"I don't have GPUs server"

"I don't have a deep leaning expert"

"I don't have labelled data" (or too few)

"I don't have the time to wait for model training "

I don't want to pay to pay for private apis" / "I'm afraid their labelling will change over time"

data iku

Solution 1 : Pre trained models

"I don't have (or few) labelled data"

-> Is there similar data ?

US

PLACES DATABASE

205 categories
2.5 M images

SUN DATABASE

307 categories
110 K images

Solution 1 : Pre trained models

If there is open data, there is an open pre trained model !

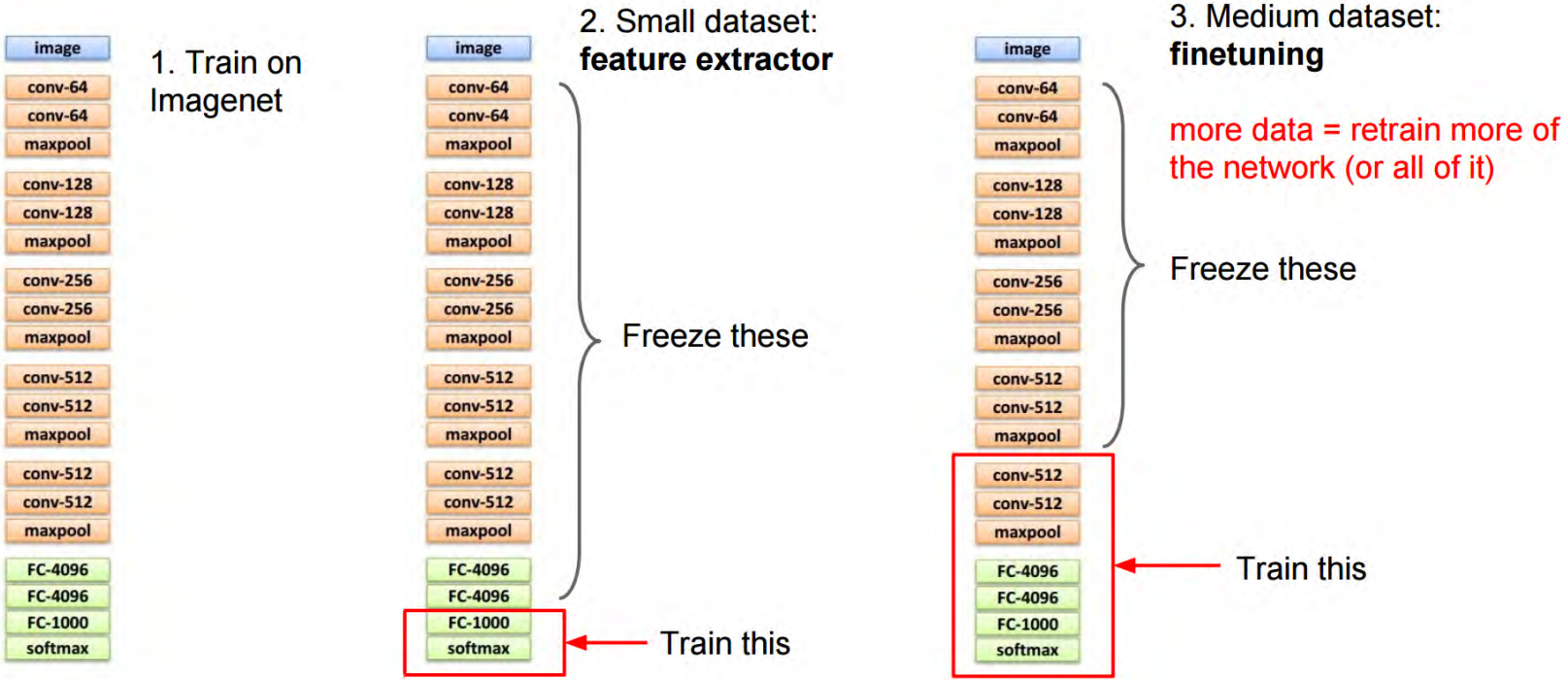
- Kudos to the community
- Check the licensing

Example with Places (Caffe Model Zoo) :

swimming_pool/outdoor: 0.65
inn/outdoor: 0.06

tower: 0.53
skyscraper: 0.26

Solution 2 : Transfer Learning



Credit : Fei-Fei Li & Andrej Karpathy & Justin Johnson http://cs231n.stanford.edu/slides/winter1516_lecture11.pdf

Solution 2 : Transfer Learning

Leverage existing knowledge !

PLACES DATABASE

SUN DATABASE

OUR DATA

Training (optional)

Caffe Model Zoo

Pre-trained model
VGG16

GPU

Transferred Data :
Last convolutional layer features

GPU

Re-Training

CPU

Re-trained model
TensorFlow
2 fully connected layers

tower: 0.53
skyscraper: 0.26

Accuracy: 72%, Top-5 Acc: 90 % > state of the art on dataset alone

data iku

Post Treatment & Results

Using Images information for BI on steroids

(Or how we transfer the labelling information)

data iku

Labels post-processing

Issue with our approach:

Complementary information

	label	proba
0	islet	0.432458
1	coast	0.198517
2	sandbar	0.164784
3	ocean	0.084271
4	sky	0.059474

Redondant information

	label	proba
0	conference_center	0.334199
1	conference_room	0.317581
2	auditorium	0.203089
3	ballroom	0.091975
4	banquet_hall	0.038570

Solution : NMF Matrix Factorization

Dimension Reduction

Sparsity

Balancedness

Explicability

data iku

Image content detection

Topic scores determine the importance of topics in an image

TOPIC	TOPIC SCORE (%)
Golf course – Fairway – Putting green	31
Hotel – Inn – Apartment building outdoor	30
Swimming pool – Lido Deck – Hot tub outdoor	22
Beach – Coast - Harbor	17

TOPIC	TOPIC SCORE (%)
Tower – Skyscraper – Office building	62
Bridge – River – Viaduct	38

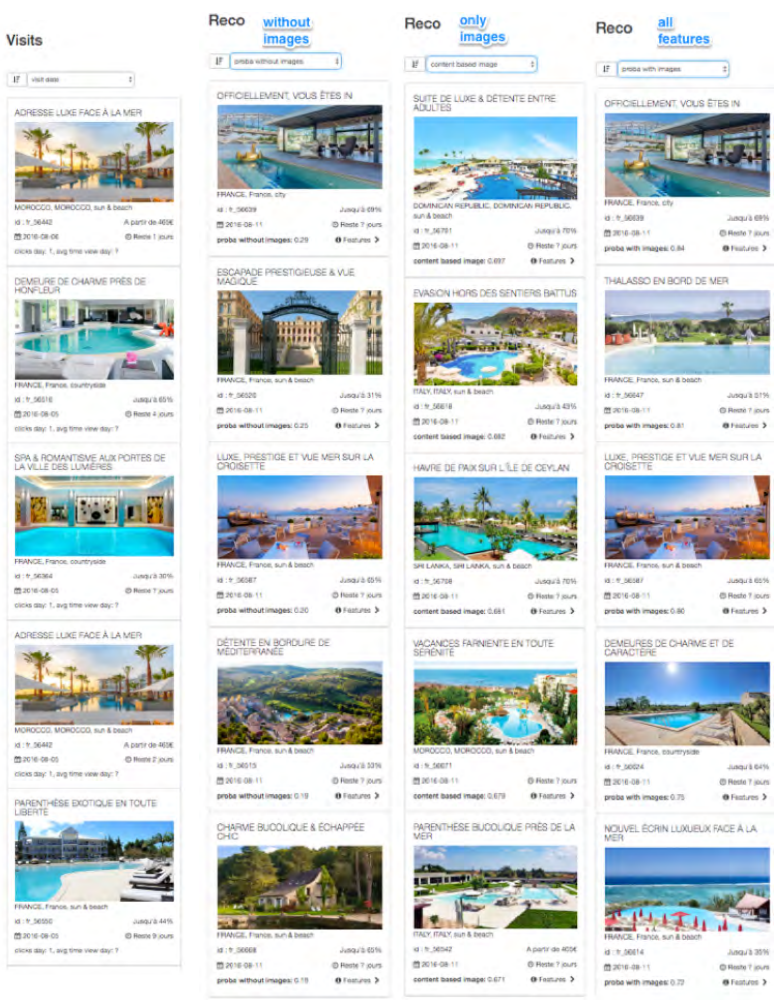
Results ?

1) Visits :
- France and Morocco
- Pool displayed

2) First Recommendation
- Mostly France & Mediterranean
- Fails to display pools

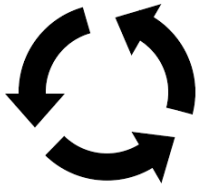
3) Only Images recommendation
- Pool all around the world
- Does not respect budget

4) Third column = Right Mix



data iku

Conclusion



Do iterative data science !

Start simple and grow
Evaluate at each steps
Image labelling = BI on steroids

Deep Learning

Don’t start from scratch !
Is there existing data ?
Is there a pre-trained model ?

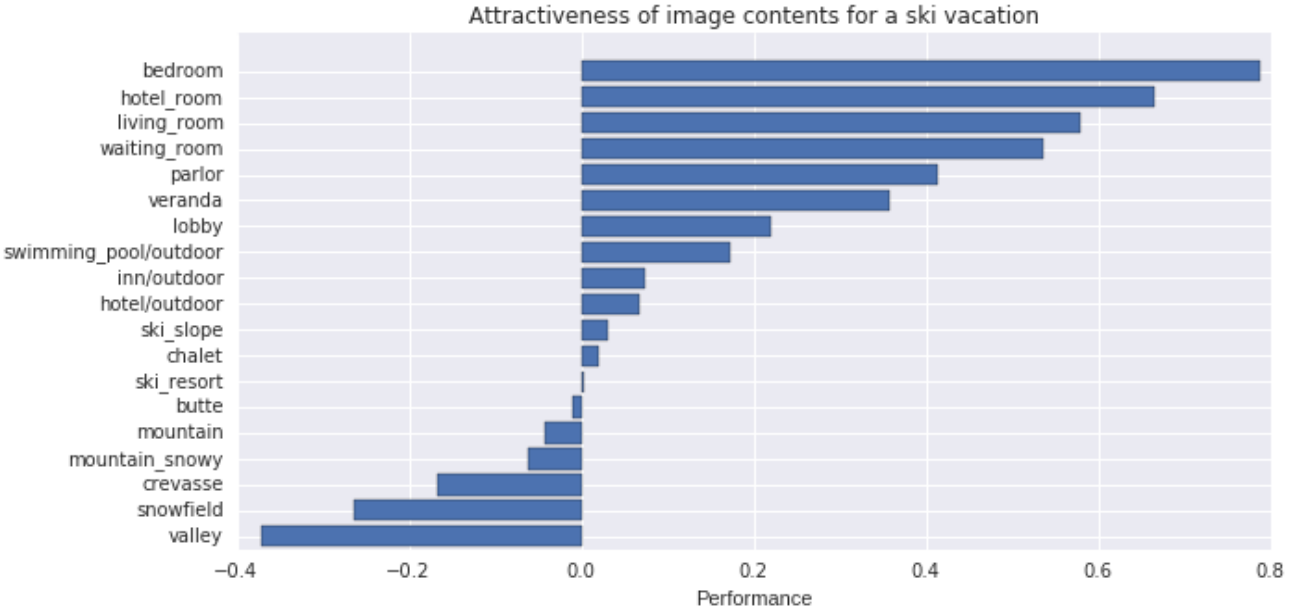
Transfer Learning

Kick-start your project
Gain time and money
Any Data Scientist can do it

What's next ?

Learned along the way

For ski sales, indoor pictures performs better



Attractiveness = % visits with tag / % sales with tag

data iku

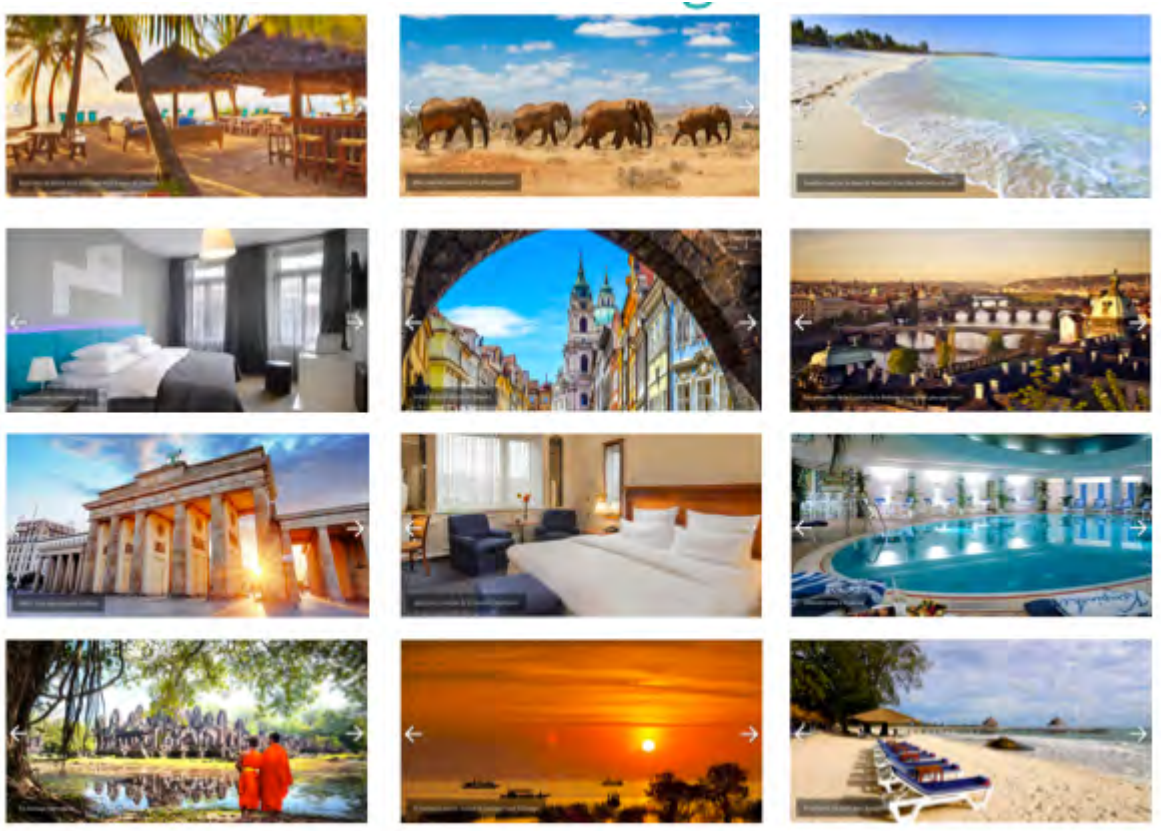
What’s Next ?

Kenya

Prague

Berlin

Cambodia



What's Next ? Customize the Image !

Kenya

Prague

Berlin

Cambodia

Thank you for your attention !

data
iku