

# Machine & Deep Learning

B&M

for

amazon

# Agenda

1. Contexte
2. Données
3. Problématiques
4. Méthodologie
5. Evaluation
6. Conclusion

## 1. Contexte

# Objectif

**Améliorer** les services de **streaming** et de **vente de produits cinéma et musique**

Grace aux données recueillies:                      ~200'000 produits (Movies & TV)  
   ~8'700'000 avis

# Environnement Technique



## 2. Données

# Exploration

### Reviews (8'765'568)

```
-- asin: string
-- image: array
-- reviewerID: string
-- reviewerName: string
-- style: struct
-- reviewTime: timestamp
-- verified: Boolean
-- vote: integer
-- summary: string
-- reviewText: string
-- overall: double
```

### Products (203'766)

```
-- also_buy: array
-- also_view: array
-- asin: string
-- brand: string
-- category: array
-- date: string
-- description: array
-- details: struct
-- feature: array
-- fit: string
-- imageURL: array
-- imageURLHighRes: array
-- main_cat: string
-- price: string
-- rank: string
-- similar_item: string
-- tech1: string
-- tech2: string
-- title: string
```

# Nettoyage

### Reviews (8'528'421)

```
-- asin: string
-- reviewerID: string
-- reviewerName: string
-- reviewTime: timestamp
-- verified: Boolean
-- vote: integer
-- summary: string
-- reviewText: string
-- overall: double
```

1 - drop duplicates

### Products (181'839)

0	asin	181'839	not null
1	title	181'781	not null
2	main_cat	181'795	not null
3	price	96'986	not null
4	description	154'977	not null
5	image	35'815	not null
6	brand	121'134	not null
7	rank_	180'002	not null
8	rank_cat	180'002	not null

1 - drop duplicates

### Priced Products (96'952)

0	asin	96'952	not null
1	title	96'952	not null
2	main_cat	96'952	not null
3	price	96'952	not null
4	description	96'952	not null
5	image	19'979	not null
6	brand	73'120	not null
7	rank_	96'713	not null
8	rank_cat	96'713	not null

1 - price non null  
2 - price < 600\$  
3 - title non null  
4 - price cleaning  
5 - cat cleaning  
6 - desc cleaning



### 3. Problématiques

## Projet 1

Analyse de sentiment sur les avis laissés par les utilisateurs d'Amazon



#### Objectifs

- Prédire la note laissée par l'utilisateur
- Identifier les critères importants pour les utilisateurs

## Projet 2

Estimation du prix du produit en fonction des méta-données

#### Objectifs

- Fournir une proposition de prix aux vendeurs Amazon



# Demo



## 4. Méthodologie

# Outils utilisés



Project storage



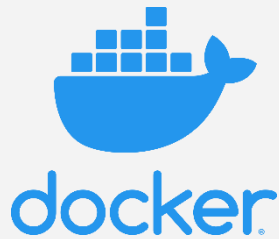
Data cleaning



Web engine



Database



Environment



Deep Learning



Main Language





# Projet 1: 5-Star rating

Choix du modèle:

`nlptown/bert-base-multilingual-uncased-sentiment`

Language	Accuracy (exact)	Accuracy (off-by-1)
English	67%	95%
Dutch	57%	93%
German	61%	94%
French	59%	94%
Italian	59%	95%
Spanish	58%	95%

Performance sur les données Amazon:

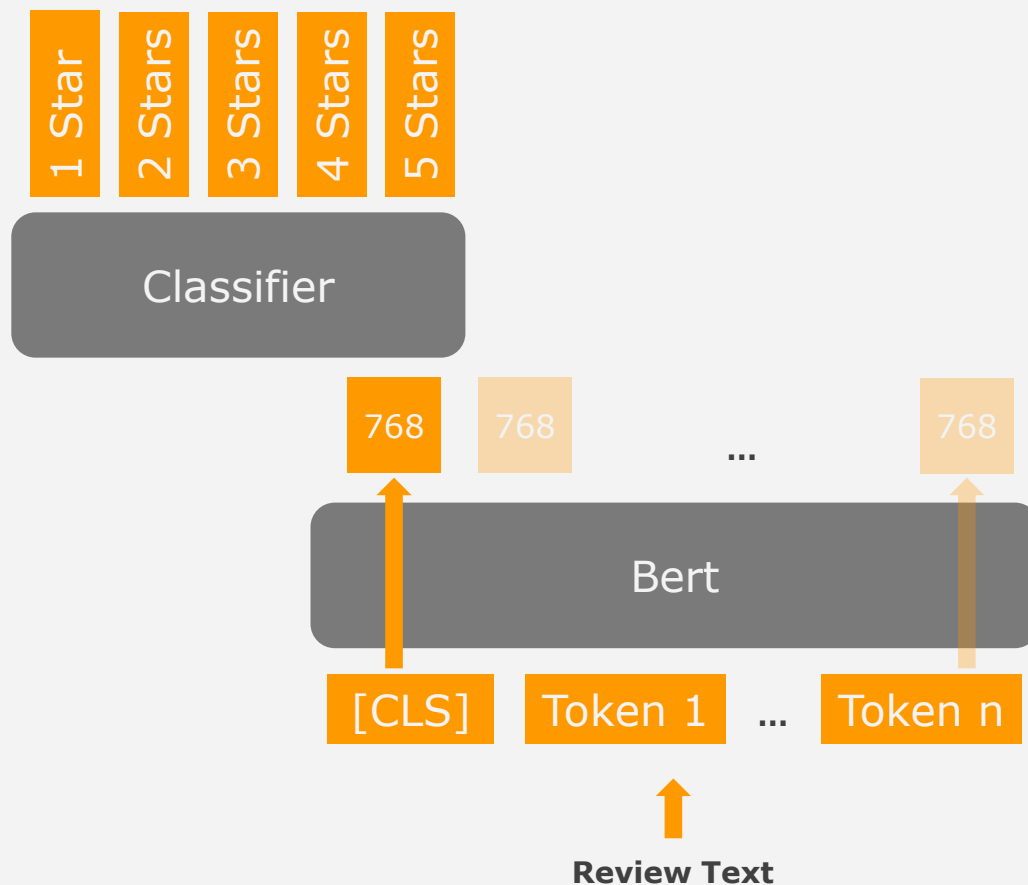
Echantillon: 4'000'000 (6h)

Language	Accuracy (exact)	Accuracy (Top2)
English	65%	87%

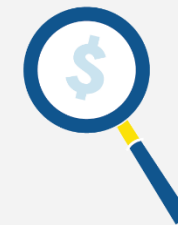




# Projet 1: 5-Star rating



## 5. Evaluation





# Projet 2: Price Estimator

Choix du modèle: ResNeXt-101 32x8d

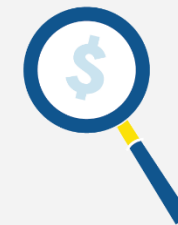
Image:

1 - FixEfficientNet-B7

FixEfficientNet-B7	87.1%	98.2%	66M	✓	Fixing the train-test resolution discrepancy: FixEfficientNet			2020
--------------------	-------	-------	-----	---	---	---	---	------

2 - ResNext

Model	#Parameters	FLOPS	Top-1 Acc.	Top-5 Acc.
ResNeXt-101 32x8d	88M	16B	82.2	96.4
ResNeXt-101 32x16d	193M	36B	84.2	97.2
ResNeXt-101 32x32d	466M	87B	85.1	97.5
ResNeXt-101 32x48d	829M	153B	85.4	97.6



# Projet 2: Price Estimator

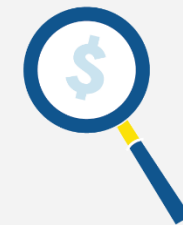
Choix du modèle:

Texte:

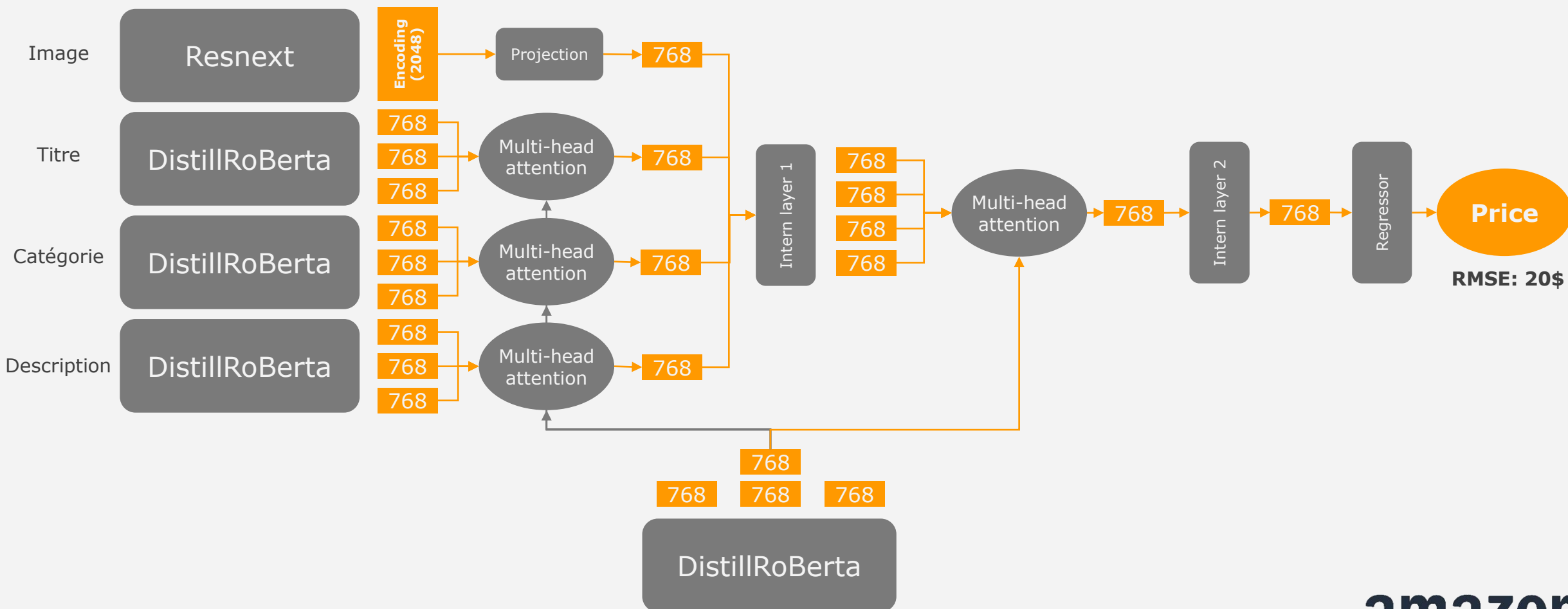
DistillRoBerta

	BERT	RoBERTa
Size (millions)	Base: 110 Large: 340	Base: 110 Large: 340
Training Time	Base: 8 x V100 x 12 days* Large: 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	Large: 1024 x V100 x 1 day; 4-5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**

## 5. Evaluation



# Projet 2: Price Estimator



## 6. Conclusion

### Projet 1: 5-Star rating



- Evaluer les critères d'importance par utilisateur
  - ➔ Affiner les suggestions faite lors de la navigation

### Projet 2: Price Estimator



- Identifier les critères permettant de mieux vendre
- Optimisation possible ?

```
Projet_AJC_Amazon
├── docker-compose.yml
├── docker_containers_config.md
├── Dockerfile
├── download_data.sh
├── env
│   ├── flask.env
│   ├── pgadmin.env
│   └── postgres.env
├── LICENSE
├── price_calculator.ipynb
├── priced_products.csv
├── README.md
├── spark_workspace
│   ├── data_ingestion.ipynb
│   ├── meta_Movies_and_TV.json
│   ├── Movies_and_TV.json
│   ├── save_as_csv.ipynb
│   └── web_app
│       ├── __init__.py
│       ├── static
│       └── templates
└── stars_model_eval.ipynb
```

**Bassem Karoui**  
**Mathias Martineau**