Supercharging GBM on GPU with Catboost

cassie.guo

This talk was given at PyDen meetup, 2020

google colab demo

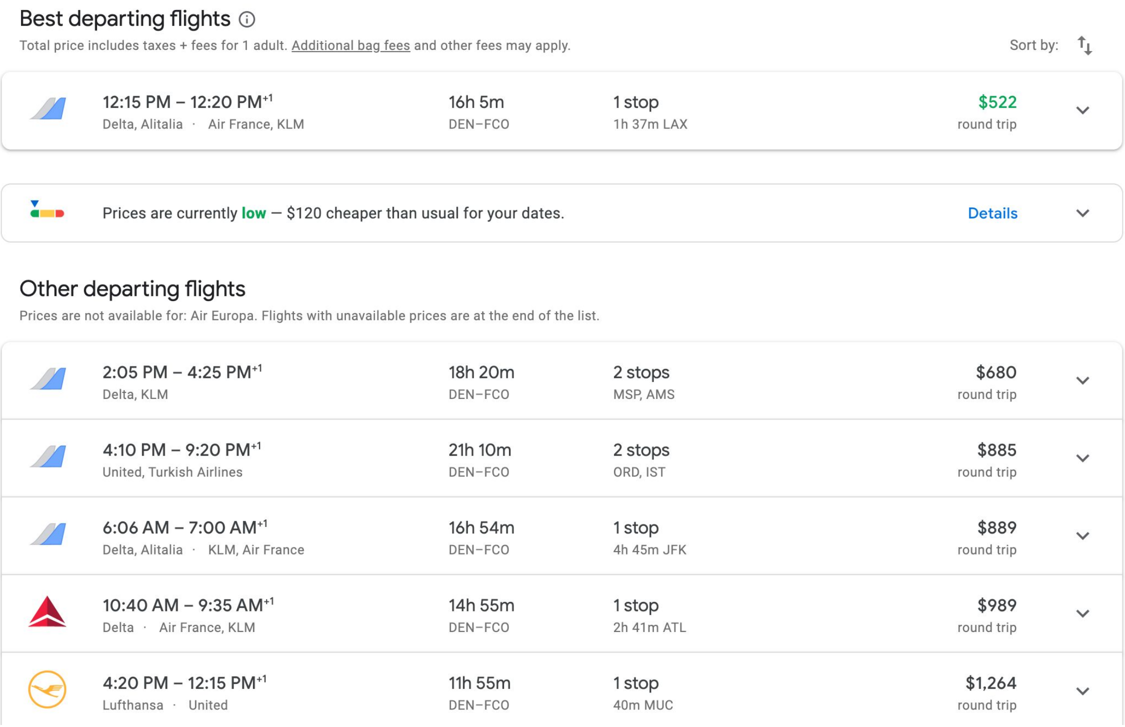
Data Scientist: The Chef Table

Data + Machine Learning → Business impact

Conversion rate (Click through rate) -- Secret sauce of targeted advertising

Facebook Hearing
Capitol Hill
3:22 PM ET
"SENATOR, WE RUN ADS"
BREAKING NEWS
LIVE
FACEBOOK CEO: "SLOW IN RECOGNIZING" RUSSIA MEDDLING THREAT
CNN
DOW ▲ 463.26
WILL "SIGNIFICANTLY LOWER" ITS 25% TARIFF ON VEHICLE IMPORTS THIS YE
NEWSROOM

Think about how google flight works...



Best departing flights ⓘ

Total price includes taxes + fees for 1 adult. Additional bag fees and other fees may apply.

Sort by: ⇅

12:15 PM – 12:20 PM^{+1} Delta, Alitalia · Air France, KLM	16h 5m DEN–FCO	1 stop 1h 37m LAX	$522 round trip

Prices are currently low — $120 cheaper than usual for your dates. Details

Other departing flights

Prices are not available for: Air Europa. Flights with unavailable prices are at the end of the list.

2:05 PM – 4:25 PM^{+1} Delta, KLM	18h 20m DEN–FCO	2 stops MSP, AMS	$680 round trip
4:10 PM – 9:20 PM^{+1} United, Turkish Airlines	21h 10m DEN–FCO	2 stops ORD, IST	$885 round trip
6:06 AM – 7:00 AM^{+1} Delta, Alitalia · KLM, Air France	16h 54m DEN–FCO	1 stop 4h 45m JFK	$889 round trip
10:40 AM – 9:35 AM^{+1} Delta · Air France, KLM	14h 55m DEN–FCO	1 stop 2h 41m ATL	$989 round trip
4:20 PM – 12:15 PM^{+1} Lufthansa · United	11h 55m DEN–FCO	1 stop 40m MUC	$1,264 round trip

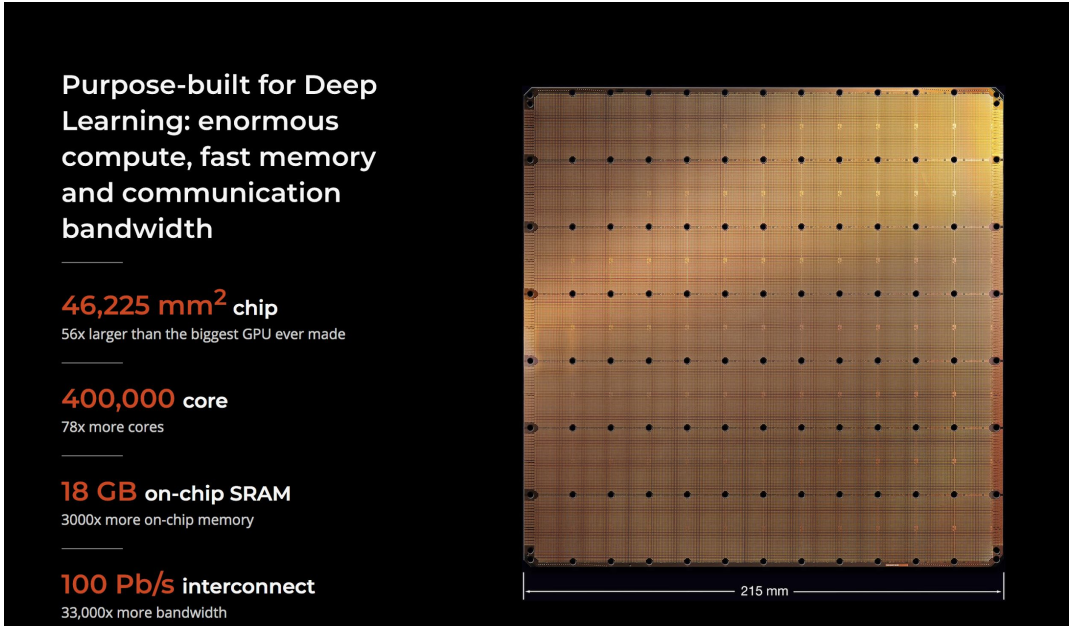
The raw data is 20-30 TB/day with 20% annual increase.

XXL data is coming

search response data

Common enterprise structured datasets

In a perfect world… More GPU, more memory, more power!
“Wafer-scale engine”



Purpose-built for Deep Learning: enormous compute, fast memory and communication bandwidth

46,225 mm^2 chip
56x larger than the biggest GPU ever made

400,000 core
78x more cores

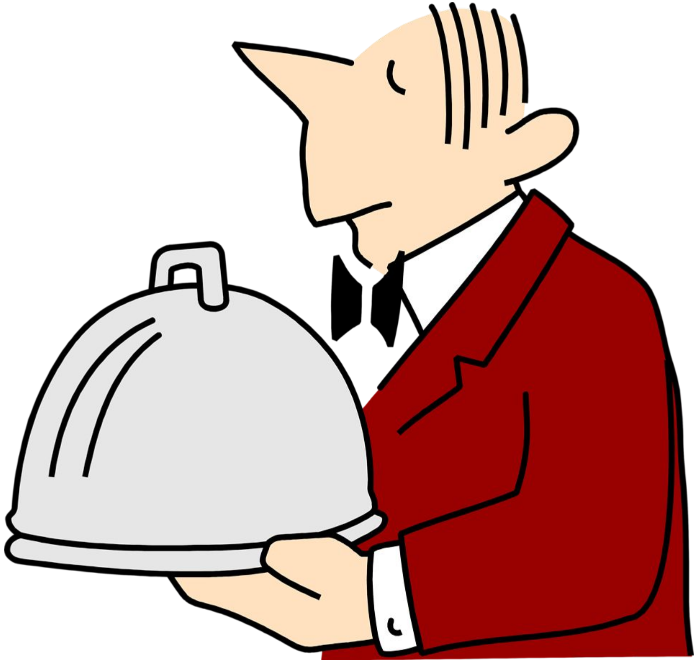
18 GB on-chip SRAM
3000x more on-chip memory

100 Pb/s interconnect
33,000x more bandwidth

source: https://www.cerebras.net/

Reality check

- 1 Tesla P100 GPU (3564 cores)
- GPU memory: 16G
- CUDA version: 9.0
- HDFS + redhat servers



Don’t reinvent the wheel

- What is the state of art model design?
- What kind of architecture to use?
- https://quinonero.net/Publications/predicting-clicks-facebook.pdf
- The scale is comparable (750 million daily active users; 1 million advertisers)
- The design is widely used and it is replicable once we figure out:
  - Feature engineering
  - Implementation of the stacking

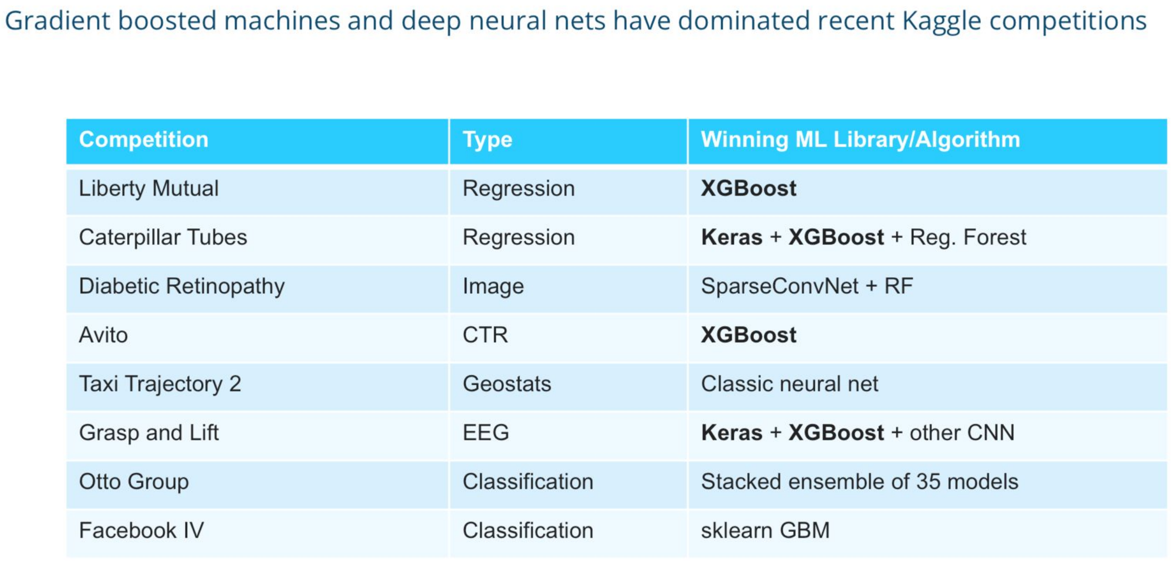
Why Catboost?

- Robust integration on GPU and multi-GPU
- Provides a variety algorithms and loss functions
- Special way of optimizing categorical data
  - Symetric trees
  - Target encoding, permutation, greedy combination

https://catboost.ai/

https://www.youtube.com/watch?v=8o0e-r0B5xQ&t=1233s

Swiss army knife for machine learning -- GBM



Gradient boosted machines and deep neural nets have dominated recent Kaggle competitions

Competition	Type	Winning ML Library/Algorithm
Liberty Mutual	Regression	**XGBoost**
Caterpillar Tubes	Regression	**Keras** + **XGBoost** + Reg. Forest
Diabetic Retinopathy	Image	SparseConvNet + RF
Avito	CTR	**XGBoost**
Taxi Trajectory 2	Geostats	Classic neural net
Grasp and Lift	EEG	**Keras** + **XGBoost** + other CNN
Otto Group	Classification	Stacked ensemble of 35 models
Facebook IV	Classification	sklearn GBM

source:
https://www.quora.com/What-machine-learning-approaches-have-won-most-Kaggle-competitions

Let’s Stacking!

https://quinonero.net/Publications/predicting-clicks-facebook.pdf

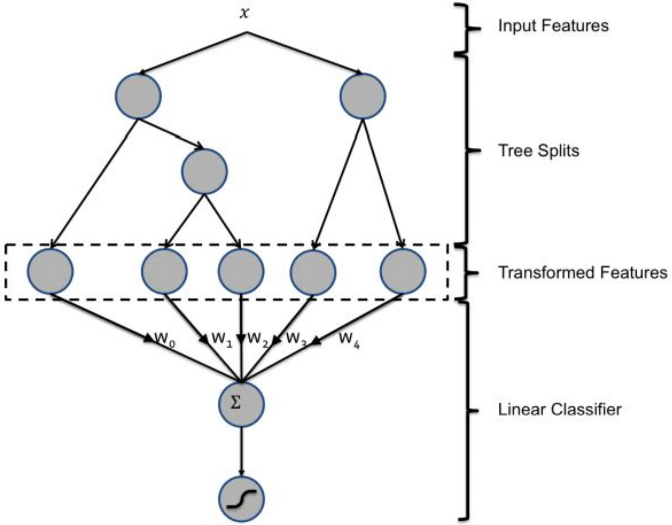
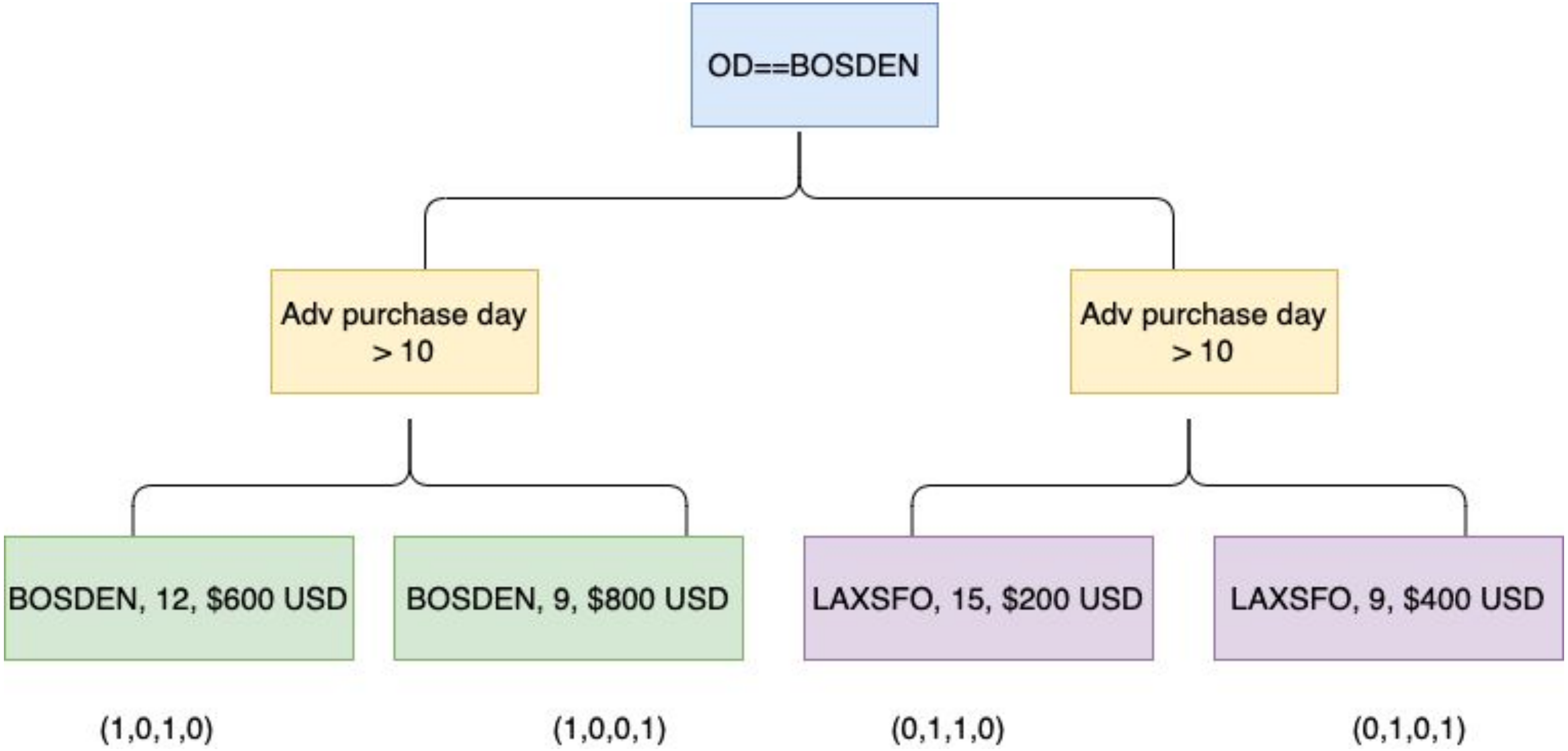


Figure 1: Hybrid model structure. Input features are transformed by means of boosted decision trees. The output of each individual tree is treated as a categorical input feature to a sparse linear classifier. Boosted decision trees prove to be very powerful feature transforms.

Vectorization of features



OD==BOSDEN

Adv purchase day > 10

Adv purchase day > 10

BOSDEN, 12, $600 USD

BOSDEN, 9, $800 USD

LAXSFO, 15, $200 USD

LAXSFO, 9, $400 USD

(1,0,1,0)

(1,0,0,1)

(0,1,1,0)

(0,1,0,1)

How we hacked them together

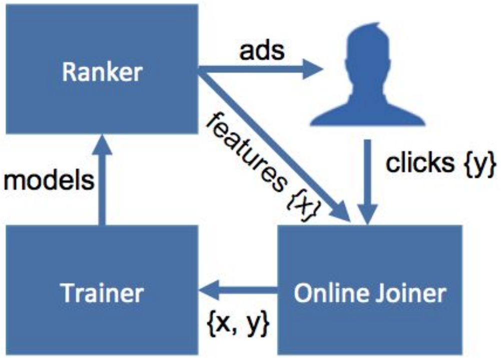
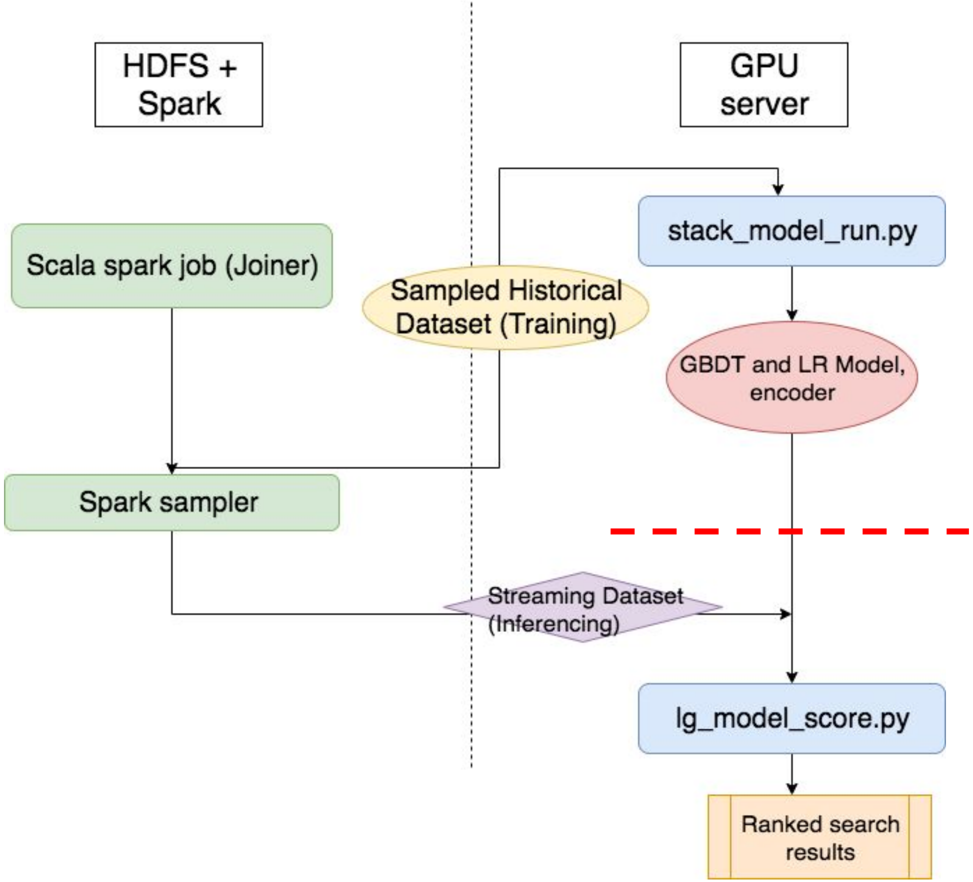


Figure 4: Online Learning Data/Model Flows.



Features are refined ingredients

- **Contextual features**
  - Market
  - Departure date
  - Advance purchase days
  - Departure hour
  - DOW, DOM
- **Historical features**
  - Past conversion rate
  - Hot markets
  - Load factor of the flight

Why it tastes good?

- Nonlinear + linear
- Convexity of the loss function (LR)
- Online + offline

Technical Debt of ML

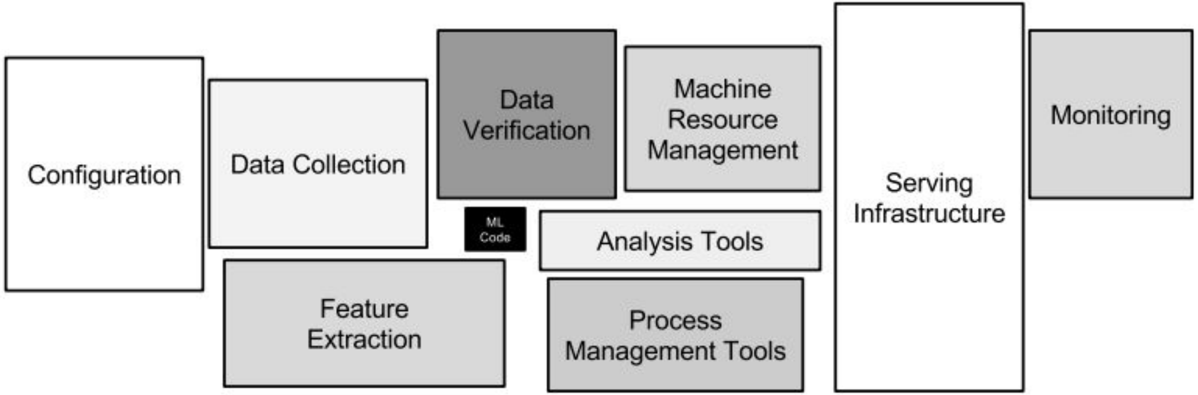


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Thank you!

Questions?