alimama · tech | DataFunSummit

2021

Deep Learning Framework Session Opening Intro

What we talk about When we talk about DL Framework

Biye Jiang Staff Engineer
Alimama Infra Efficiency team

Bio of myself

Visualization Tools

How to build interactive viz tools?

- imMens (Database in WebGL)
- Data illustrator @ Adobe Intern

BIDMach

How to make ML easy to use?

- ML Framework written in Scala
- Interactive ML @ Azure ML Intern

XDL & Bernoulli

How to apply ML in production?

- Algorithm-System co-design
- Distributed training/serving
- Resource allocation under SLO

PhD in UC Berkeley

Alibaba Ads Team

BIDMach: An unusual ML framework

BiD Berkeley institute of design

BIDMach: A Scala-written ML framework, developed by John Canny & Huasha Zhao in 2012

Scala: Great tools to make DSL

- Unified structure with compact codebase
- Flexible as C++, Maintainable as Java, Interactive as Python

BIDMACH

```
def eStep(sdata:Mat, user:Mat):Unit = {
   for (i <- 0 until opts.uiter) {
     val preds = SDDMM(mm, user, sdata)
     val unew = user o (mm * (sdata / preds)) + opts.alpha
     user <-- exppsi(unew)
   }
}
```

LDA e-step code example

https://github.com/BIDData/BIDMach

Big Data Analytics with Small Footprint: Squaring the Cloud. Canny et al. KDD 2013
Interactive Machine Learning via a GPU-accelerated Toolkit. Jiang et al. IUI 2017

BIDMach Architecture

Jupyter notebook
& Interactive Visualization tools

BIDMach: ML engine
Autograd for deep learning

BIDMat
Matrix layer with caching

Sparse Allreduce

CPU
MKL/BLAS

GPU
cuBLAS/cuDNN
Customized kernel

Actor framework

People

Algorithm

System

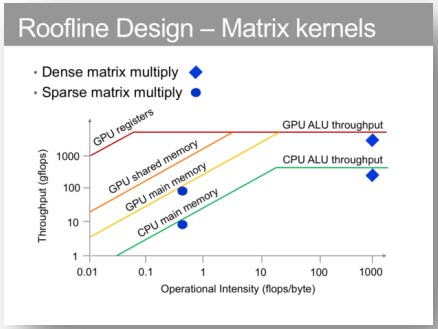
Notebook

Interactive

Parameter tuning

Visualization

Co-design towards hardware limit as well as usability

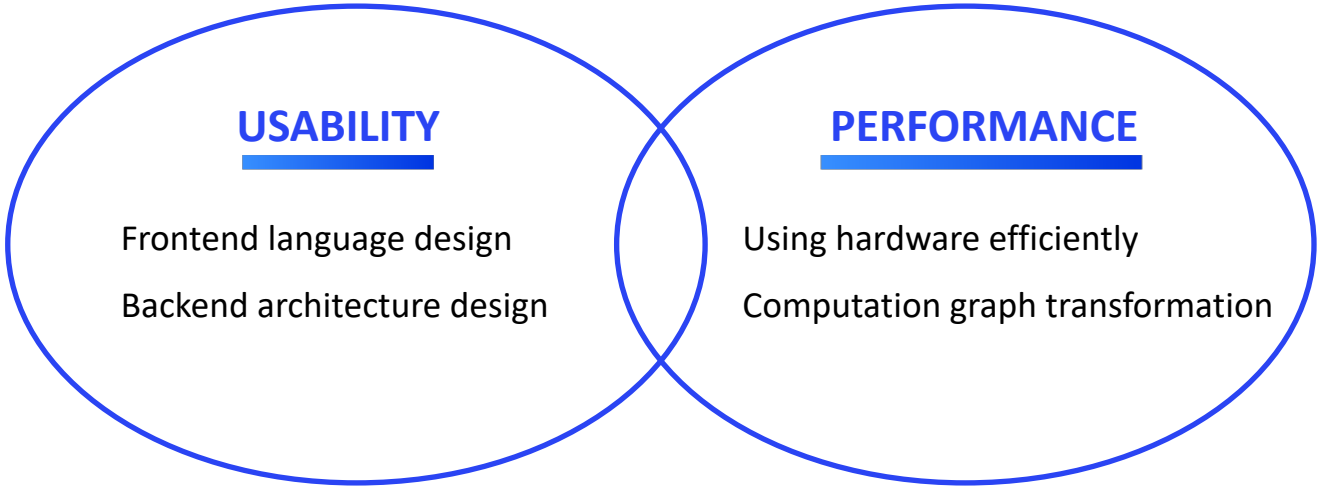


Kylix: A sparse allreduce for commodity clusters. Zhao et al. ICPP 2014
Extending the Limits of Machine Learning with GPUs @GTC2015
Interactive Machine Learning via a GPU-accelerated Toolkit. Jiang et al. IUI 2017

So far

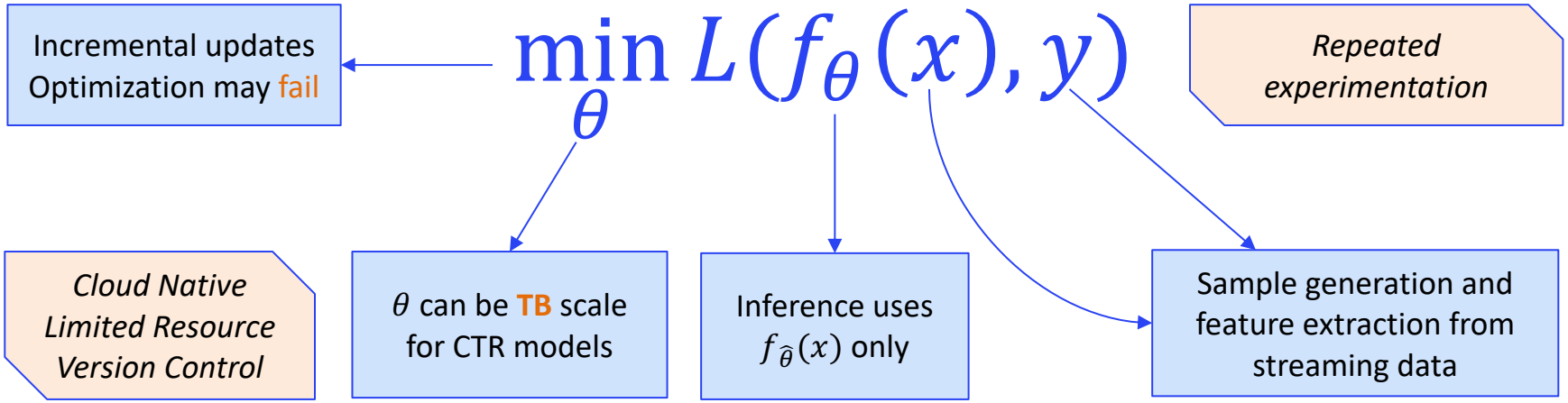
When we are talking about ML framework, we're talking about (at least):

"A piece of code to solve my problems!!!"

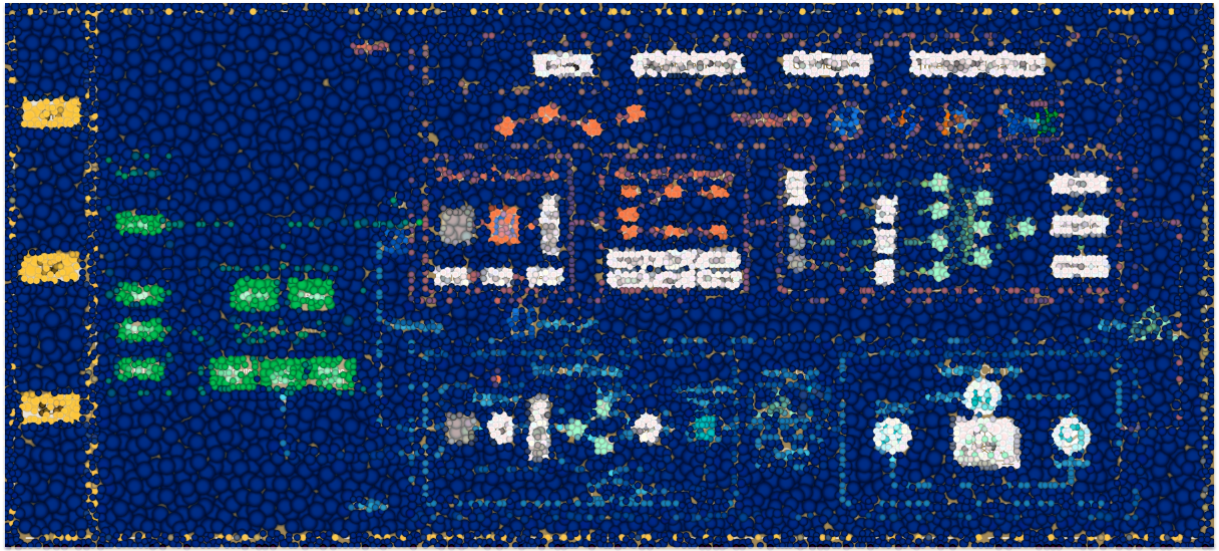


Moving into Production

- Frameworks like Tensorflow/PyTorch work very well in research
- But how does a “Framework” in Internet industry (Search Engine/Recommendation/Ads) look like?



Industrial MLSys@Alimama



How to maintain and optimize such system?

- Co-Design
- Evolution Cycles

Further Talks

- XDL -> *This afternoon*
- Bernoulli (Streaming ML) -> *Flink Forward Asia 2021*

https://github.com/alibaba/x-deeplearning

XDL: An Industrial Deep Learning Framework for High-dimensional Sparse Data. Jiang et al. DLP-KDD 2019

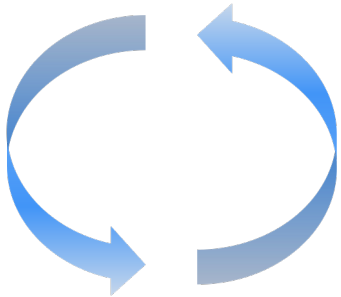
What Do We Need for Industrial Machine Learning Systems? Bernoulli, A Streaming System with Structured Designs. Luo et al. DLP-KDD 2021

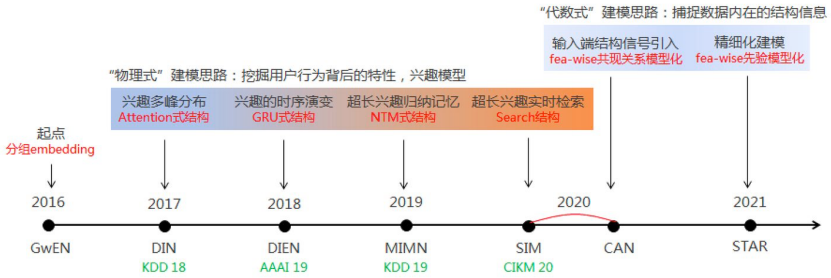
《屠龙少年与龙：漫谈深度学习驱动的广告推荐技术发展周期》2021

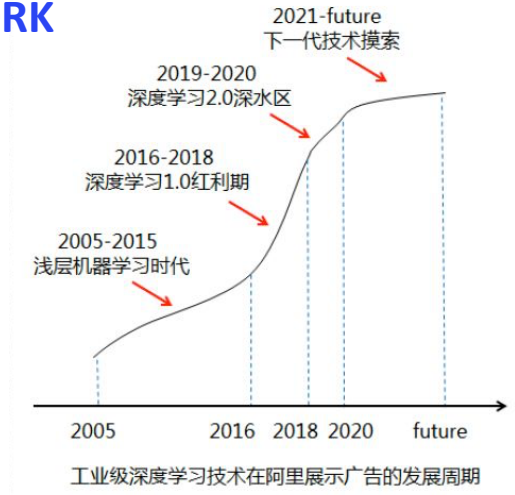
Framework – Algorithm Co-Evolution

- *"The Hardware Lottery"*

  A research idea wins because it is suited to the available software and hardware



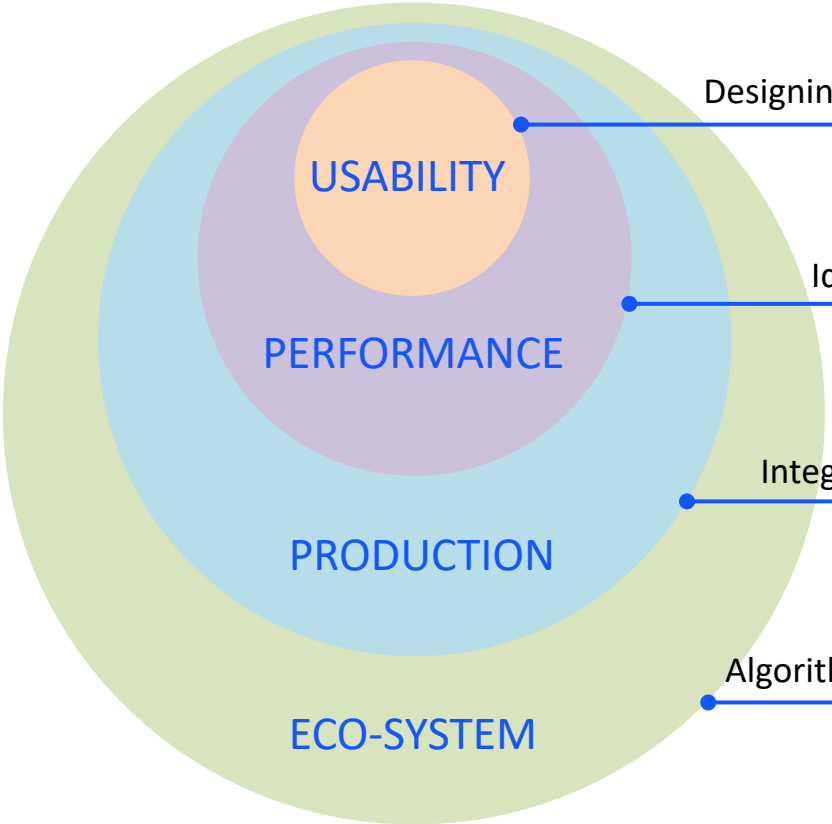




The Hardware Lottery. Sara Hooker 2020

《屠龙少年与龙：漫谈深度学习驱动的广告推荐技术发展周期》2021

Summary



Designing “Programming language”

Programming efficiency

Ideas from PL / DB / System

Machine efficiency

Integrity / Scalable / Extendable

Organization efficiency

Algorithm Evolution / Open Source

Community efficiency

Summary

- Programming languages are evolving and require designers to make tradeoffs, so do ML frameworks

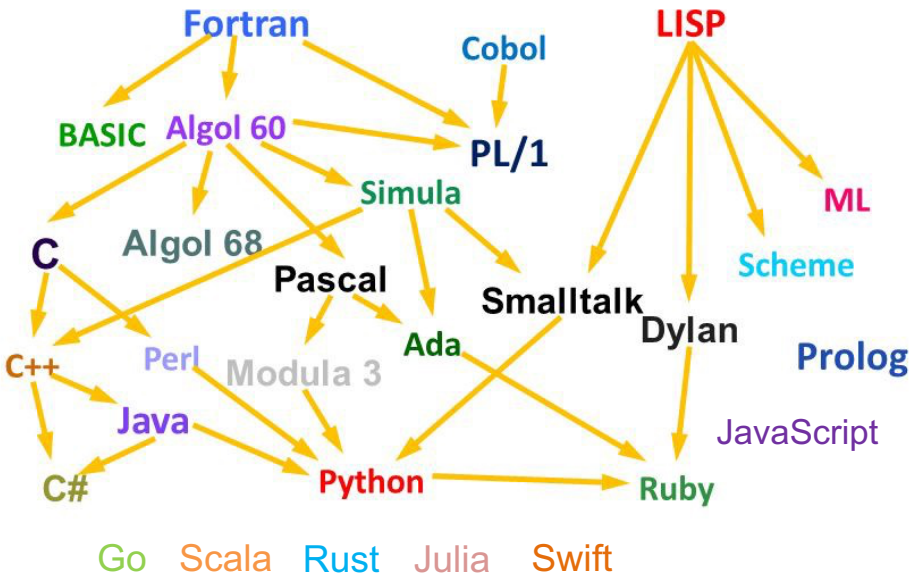


Image by Ana Harris

Caffe2 PYTORCH theano

Keras TensorFlow dmlc mxnet XDL

BIDMACH [M]S PaddlePaddle MindSpore MegEngine

ONEFLOW Jittor 计图 DGL DEEP GRAPH LIBRARY

alimama · tech | DataFunSummit

论坛议程

今天有6场精彩演讲，涵盖性能优化、框架整体设计、大模型训练、图神经网络等几大方向

主题	机构	演讲嘉宾
面向云计算的分布式机器学习优化实践	Google	蓝昶
超大模型的技术挑战与Mindspore的解决思路	华为	姚逸璠
超大模型高效训练的分布式框架Whale	阿里云	王林
飞桨：源于产业实践的开源深度学习平台	百度	蓝翔
MegEngine DTR 技术与训练框架技术创新	旷视科技	许欣然
构建高效易用的图深度学习平台	亚马逊	王敏捷

THANKS!

Contact:
biye.jby@alibaba-inc.com