MLOps on the Edge

About difference-engine.ai

difference-engine.ai is a technology consulting firm specialising in solving business problems through Data Science, Applied Machine Learning and AI Driven Products.

We are a team of 13 ML Engineers.

About Me

Prathamesh Sarang

- 6+ years experience in Software Engineering and Data Science
- Ex-Systems Engineer at Infosys
- Ex-Data Scientist at Lemoxo Technologies and Damco (Part of AP Moller Maersk Group)
- Currently working as Machine Learning Engineer (Data Products) at difference-engine.ai
- Working at the intersection of Data Science and Software Engineering

What are the interesting things I'm doing?

- Not much traditional Machine Learning these days

What are the interesting things I'm doing?

- Not much traditional Machine Learning these days
- Working on Web applications

What are the interesting things I'm doing?

- Not much traditional Machine Learning these days
- Working on Web applications
- Majority my work is into engineering around ML applications 🤓

Outside of work, what I do!

- Big history, horror and true crime fan, generally listen to podcasts, watch movies/series and read books

Outside of work, what I do!

- Big history, horror and true crime fan, generally listen to podcasts, watch movies/series and read books
- I teach as well

What can you expect from this talk

I have a few questions

How many Technical folks?

What’s the exposure of the crowd wrt ML?

How many have done ML Deployments?

What is Data Science, Machine Learning and AI?

Mat Velloso
@matvelloso

Follow

Difference between machine learning and AI:

If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

5:25 PM - 22 Nov 2018

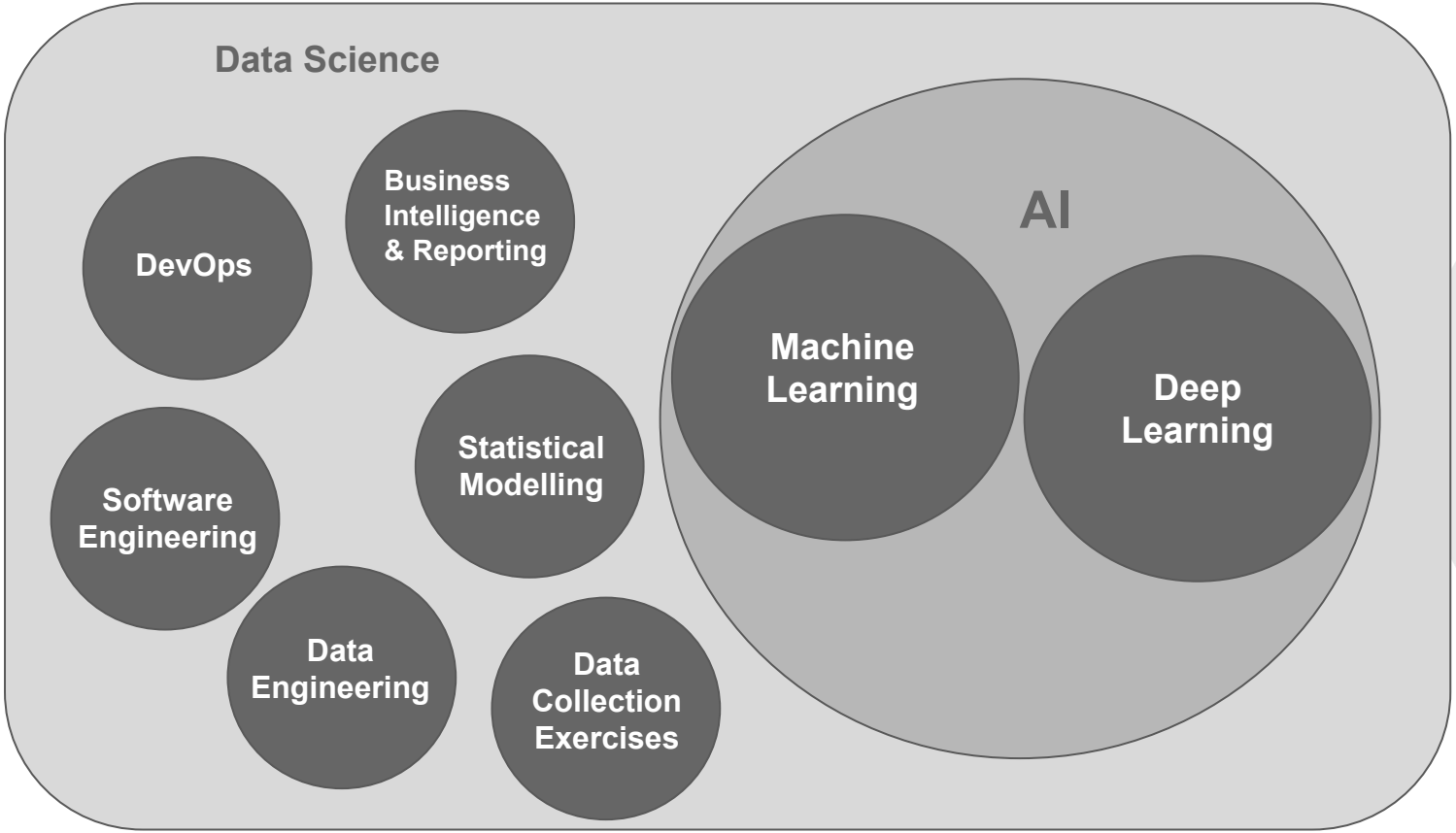
7,622 Retweets **21,608** Likes

187 7.6K 22K

difference-engine.ai

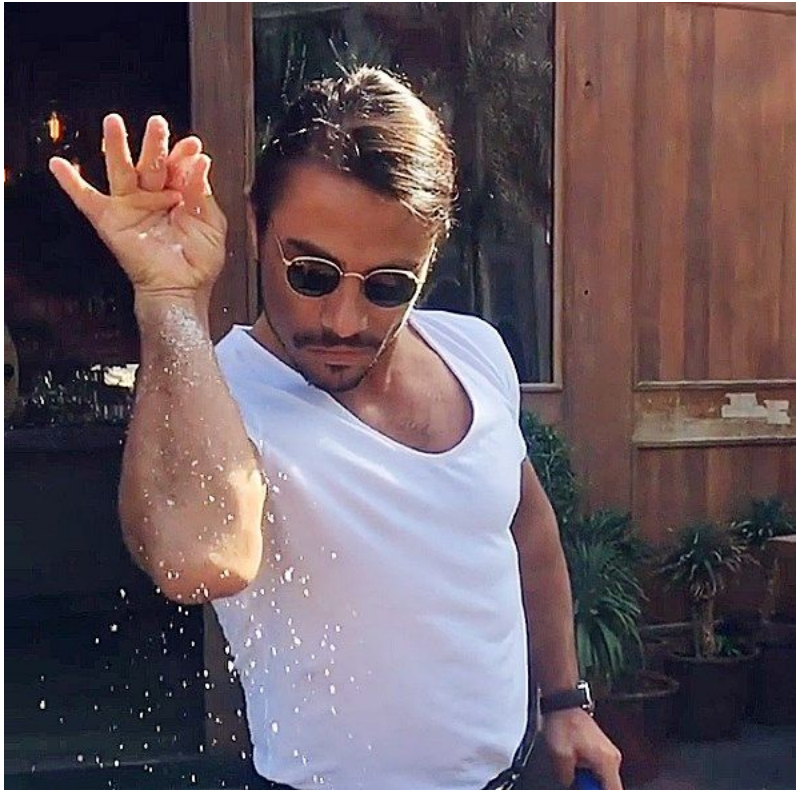
So many conflicting definitions!

I’ll add one more!



difference-engine.ai

Take that with a pinch of salt



difference-engine.ai

Google

Google Search

Google Maps

Google Translate

Google Lens

NETFLIX

Content Recommendations for users

Artwork Personalization at Netflix

Data Science and the Art of Producing Entertainment at Netflix

difference-engine.ai

Contracting and Procurement

Downstream Retail

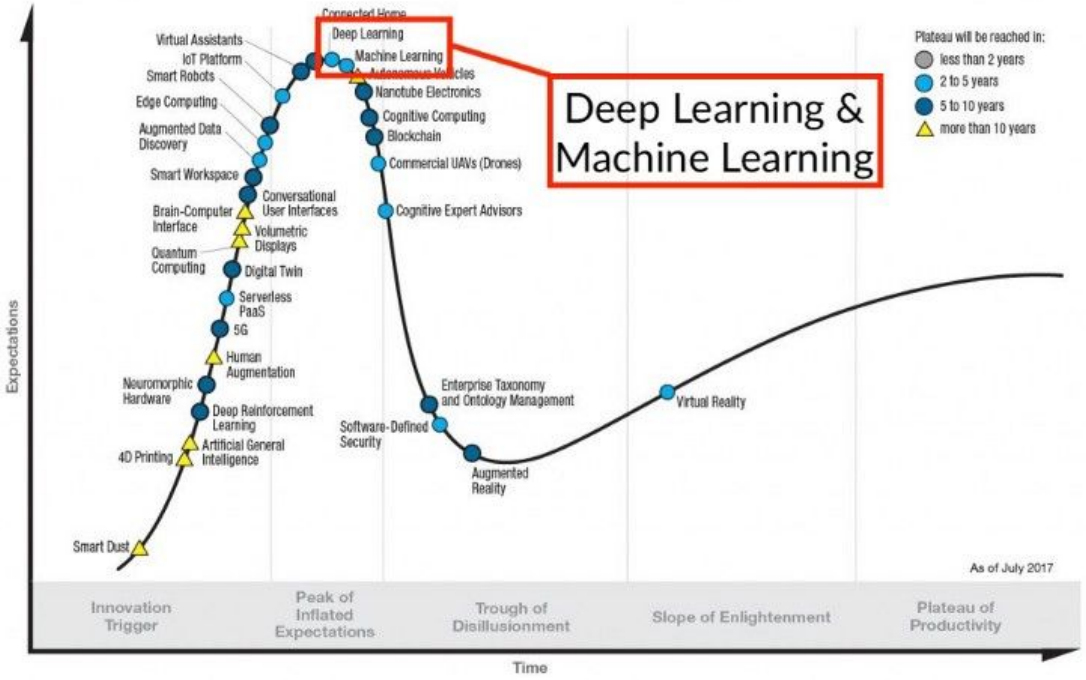
Shell Exploration

difference-engine.ai

Closer home!

Machine Learning & Deep Learning hype

Gartner Hype Cycle for Emerging Technologies, 2017

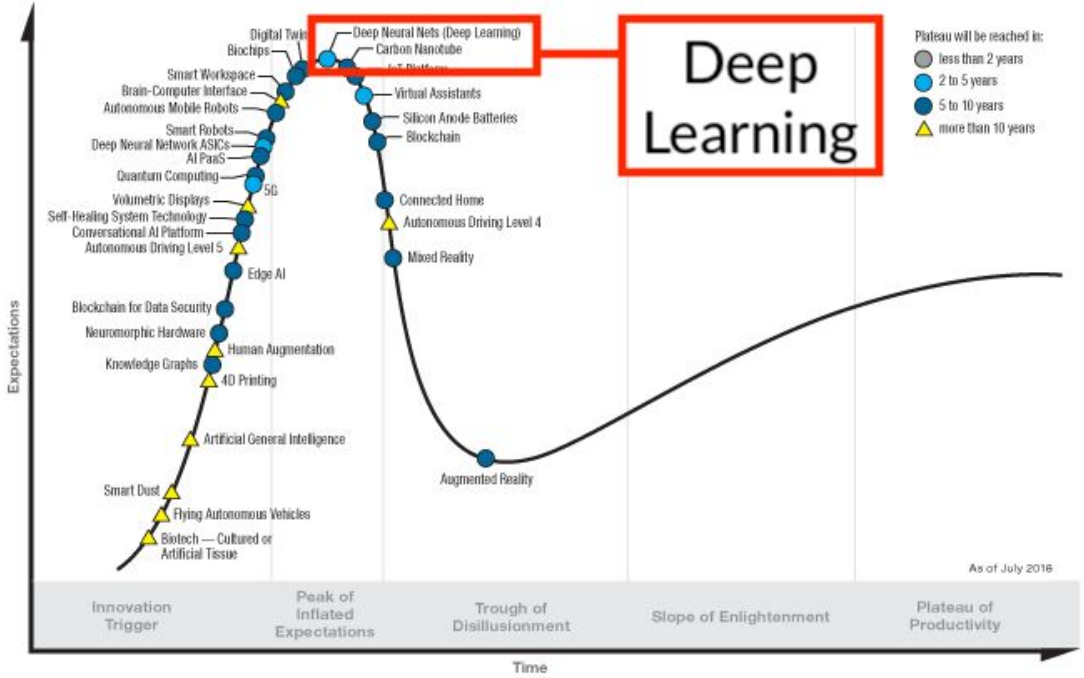


gartner.com/SmarterWithGartner

Source: Gartner (July 2017)

Gartner

Hype Cycle for Emerging Technologies, 2018



gartner.com/SmarterWithGartner

Source: Gartner (August 2018)

Gartner.

Why all of this?

To solve business problems!

But ML isn’t the most important part!

Data Collection

Feature Engineering

Checking availability of training resources on local/cloud

Formulating the problem

Managing Different Data Sources

ML/DL

Serving ML/DL Models

Error Logging, Model Health/Decay in production

Versioning Data, Code and Models run during experiments

Making ML work with existing/new Software application

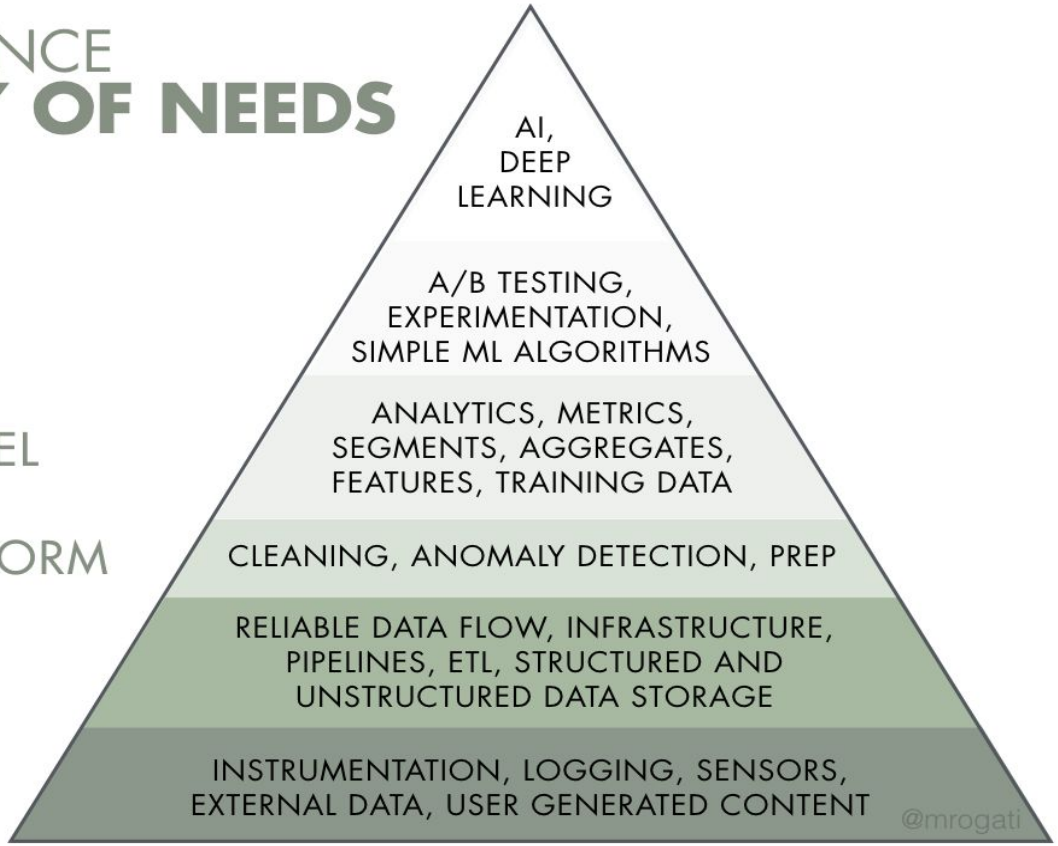
Reporting metrics/KPIs to business stakeholders

difference-engine.ai

It’s already hard!

difference-engine.ai

THE DATA SCIENCE **HIERARCHY OF NEEDS**



Source: Monica Rogati's AI Needs of Hierarchy

difference-engine.ai

Mantra to do Data Science in an organization

(Shamelessly copied from Dr. D.J. Patil)

1x

Prototyping Phase: Picking up problems and solving them

Just letting the business know that we are capable

10x

Real Deployment Strategy
comes into play

100x

In simple words, this is pure scale! Organization wide ML Adoption

0 to 1x is cool!

Let’s assume you were able to get a pretty compelling AUC 0.834

Or the Deep Learning architecture you stumbled on is great!

Your Data Scientist be like!

difference-engine.ai

While deployment!

difference-engine.ai

1x to 10x is hard for a lot of people

Friction between Software/Ops and Data Science

Code or algorithms don’t make ML hard, people do!

1x solutions

1x Solution

Dashboards

Jupyter Notebooks

Powerpoint Presentation

Problems with 1x solution

Ease of Use

Scaling it to actual users

Are you really serious about doing ML?

10x solution

10x Solutions

Dashboards

Data Engineering

APIs

On Device ML

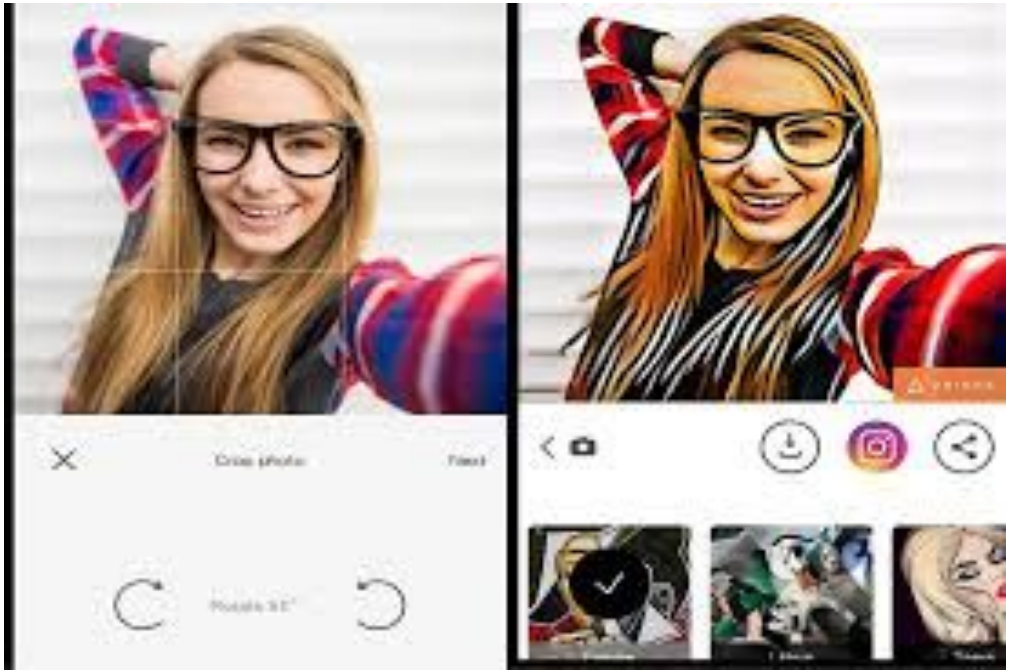
Where 10x solutions might fail

Latency

Cost per inference

Security & Privacy

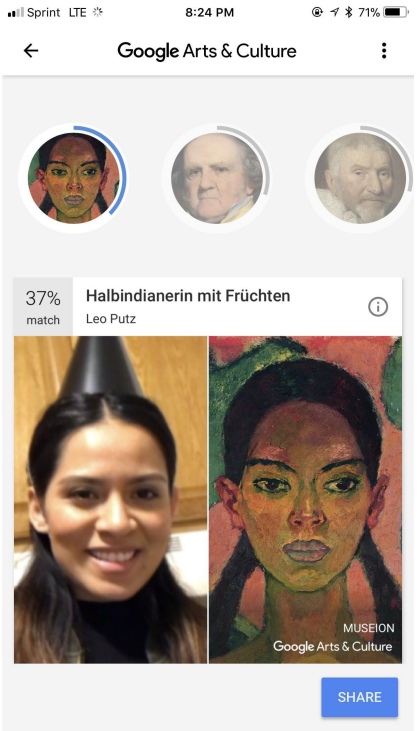
Prisma



difference-engine.ai

Prisma suffered from Latency when it started off

Google Arts and Culture App



difference-engine.ai

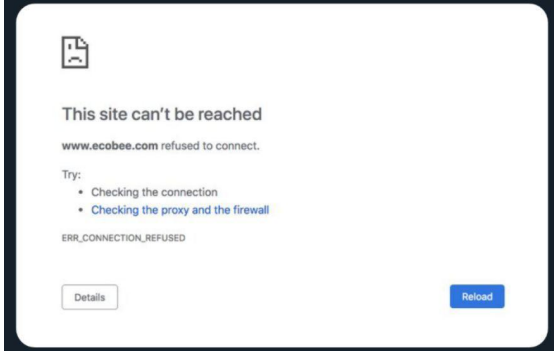
Big backlash on Social Media due to the intrusive nature of the application

IoT/Edge devices

Why am I awakened by a freezing house in 14 degree weather? Furnace is working. Vents are well maintained.

Oh, I see why.

The goddamn @ecobee server for the networked thermostat is down. And what could go wrong, in depending on the internet to keep the child warm?

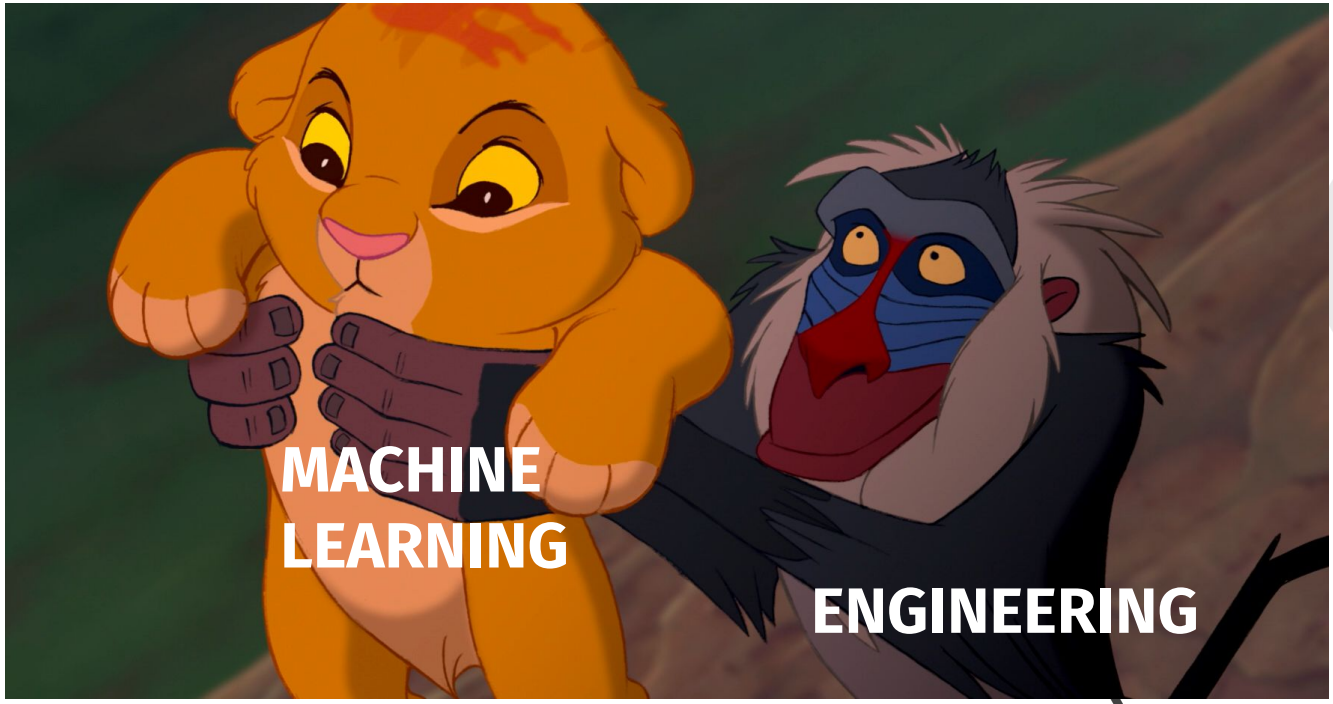


difference-engine.ai

Right now everyone is lingering around this area

100x

100x Solution



difference-engine.ai

Why are we all here?

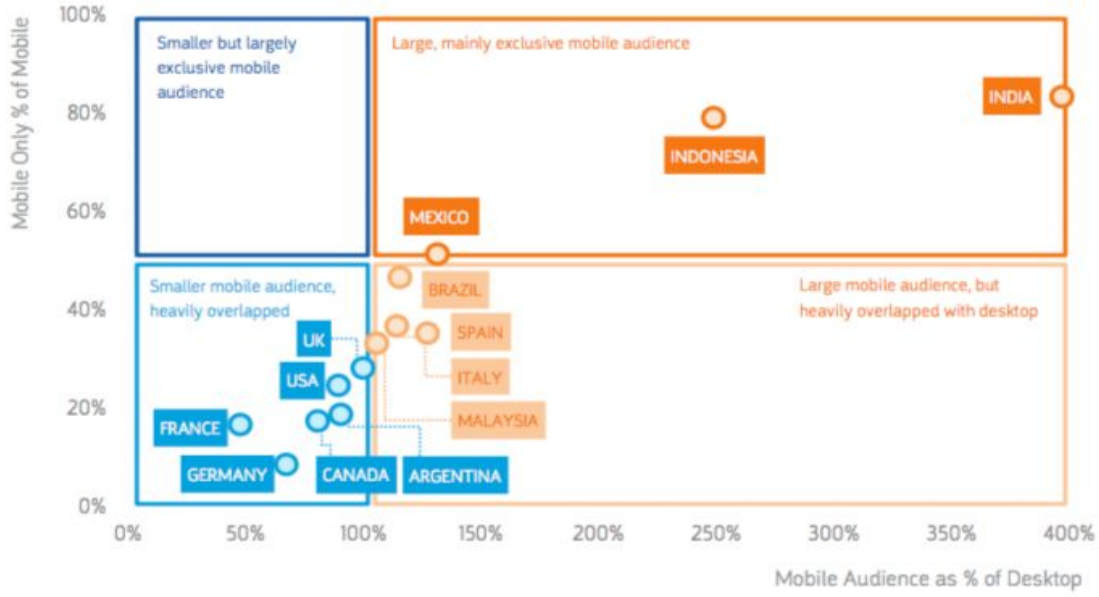
difference-engine.ai

ML on IoT/Edge devices is a great use case

'Edge' refers to the computing infrastructure that exists close to the sources of data

Can you name some edge devices?

Impact of mobile devices isn't to be ignored



comScore

difference-engine.ai

Current state for OnDevice ML

ML Kit

TensorRT

ONNX

TensorFlow Lite

difference-engine.ai

Memory constraints on device

Models are huge!

Trade Offs: Metrics vs Usability

No platform agnostic solutions

Better workflows for mobile deployments

TensorFlow

OR

Keras

+ TensorFlow Lite + Flutter

difference-engine.ai

Demo

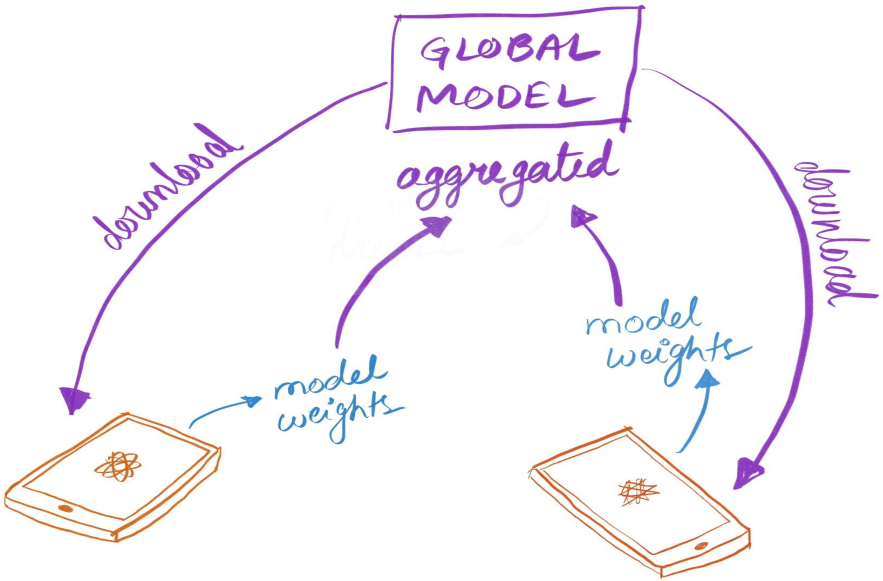
Better workflows for client-side deployments

TensorFlow.js

ONNX

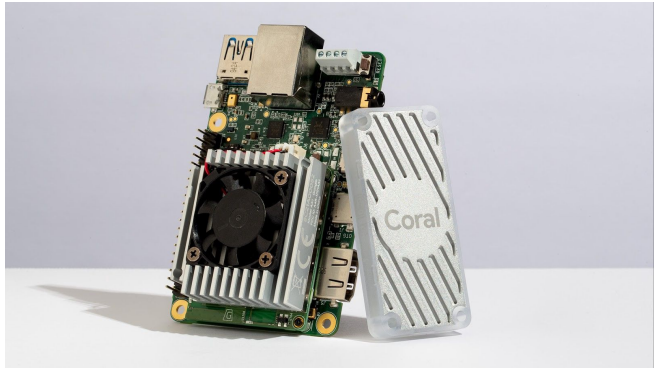
Recent developments

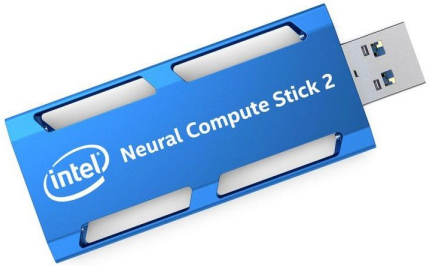
Federated Learning



difference-engine.ai

Custom Hardware to do on-device ML





Improved Model Compression techniques

Quantization

Pruning

Knowledge Distillation

Low Rank approximation

Improved Model Architectures

Where can I read?

Recent Advances in Efficient Computation of Deep Convolutional Neural Networks:
https://arxiv.org/pdf/1802.00939.pdf

Awesome Model compression and abstraction:
https://github.com/memoiry/Awesome-model-compression-and-acceleration/blob/master/README.md

Model Compression Papers: https://paperswithcode.com/task/model-compression

Thank You

Catching hold of me

Email: prathamesh.b.sarang@gmail.com

Shameless promotion

LinkedIn: Prathamesh Sarang
Personal blog: https://pratos.github.io
Twitter: prat0s

difference-engine.ai