Semi-supervised learning with GANs : putting together the pieces together

OLX Berlin Data Science Team

Classifieds


CLASSIFIEDS

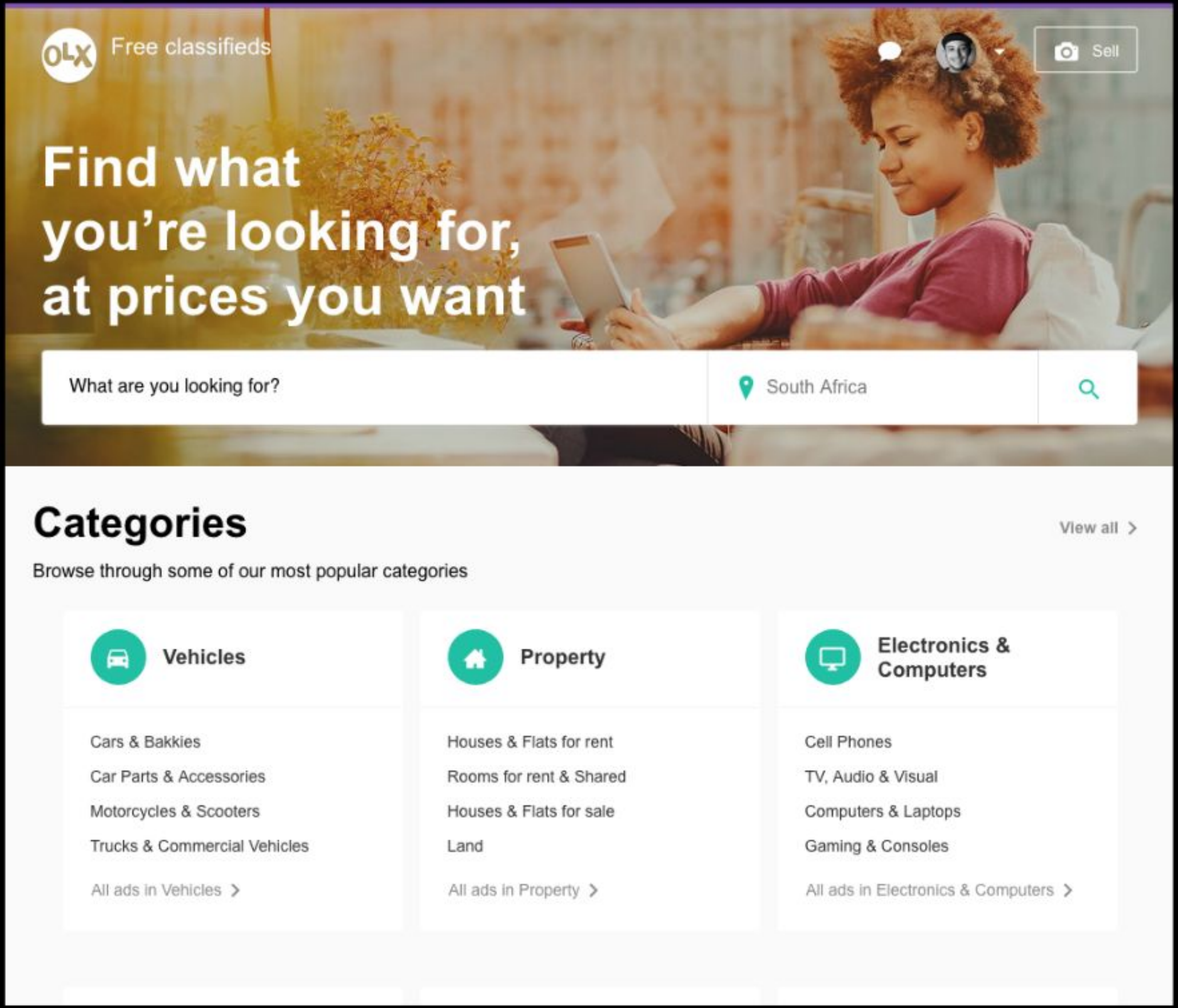
PECAN SQUARE APARTMENTS

(318) 872-2348

LENNOX

Jon Clemons Real Estate

Offline



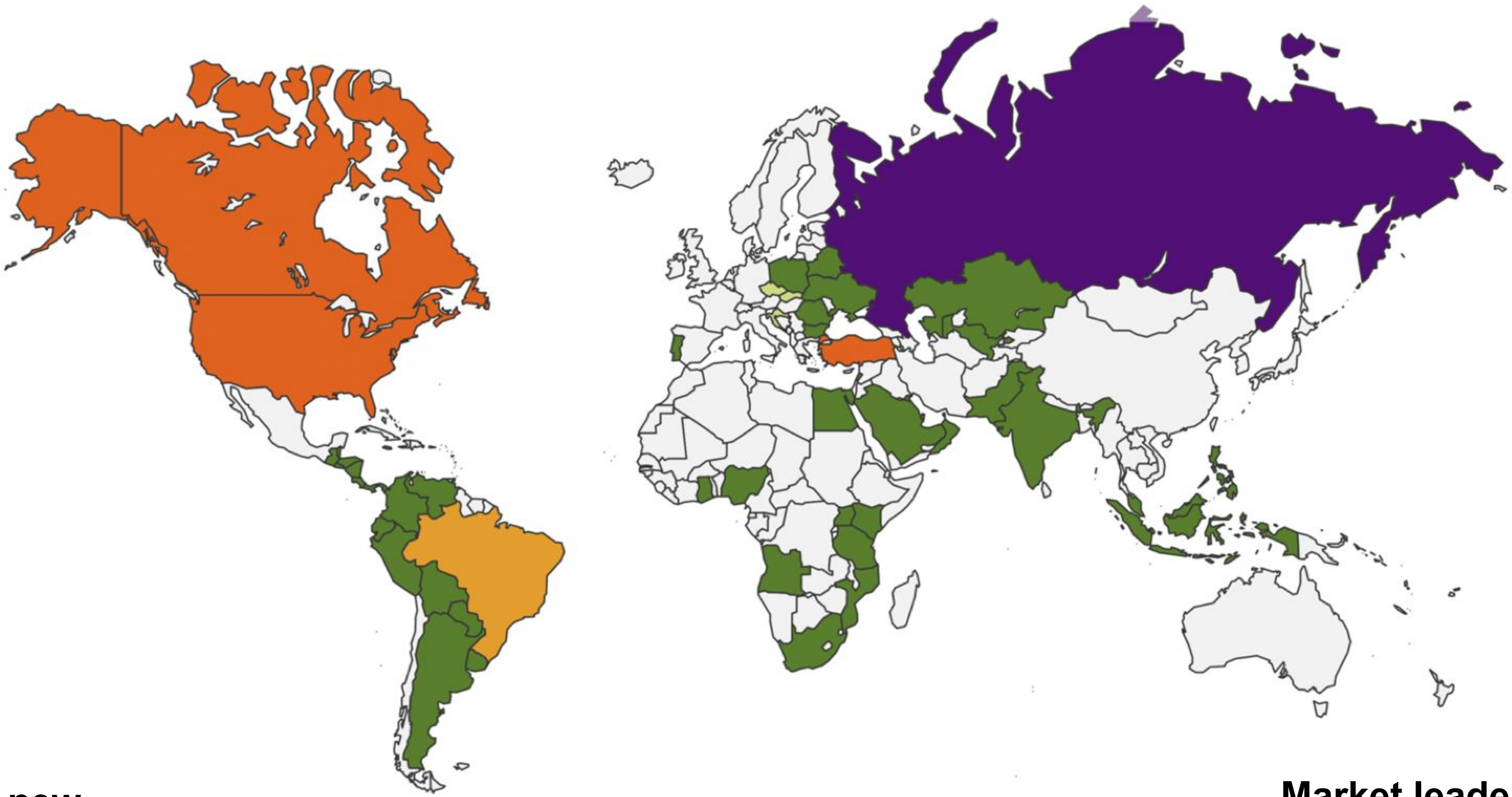
OLX

OLX GROUP

OLX in a glance

350+ million active users

40+ countries

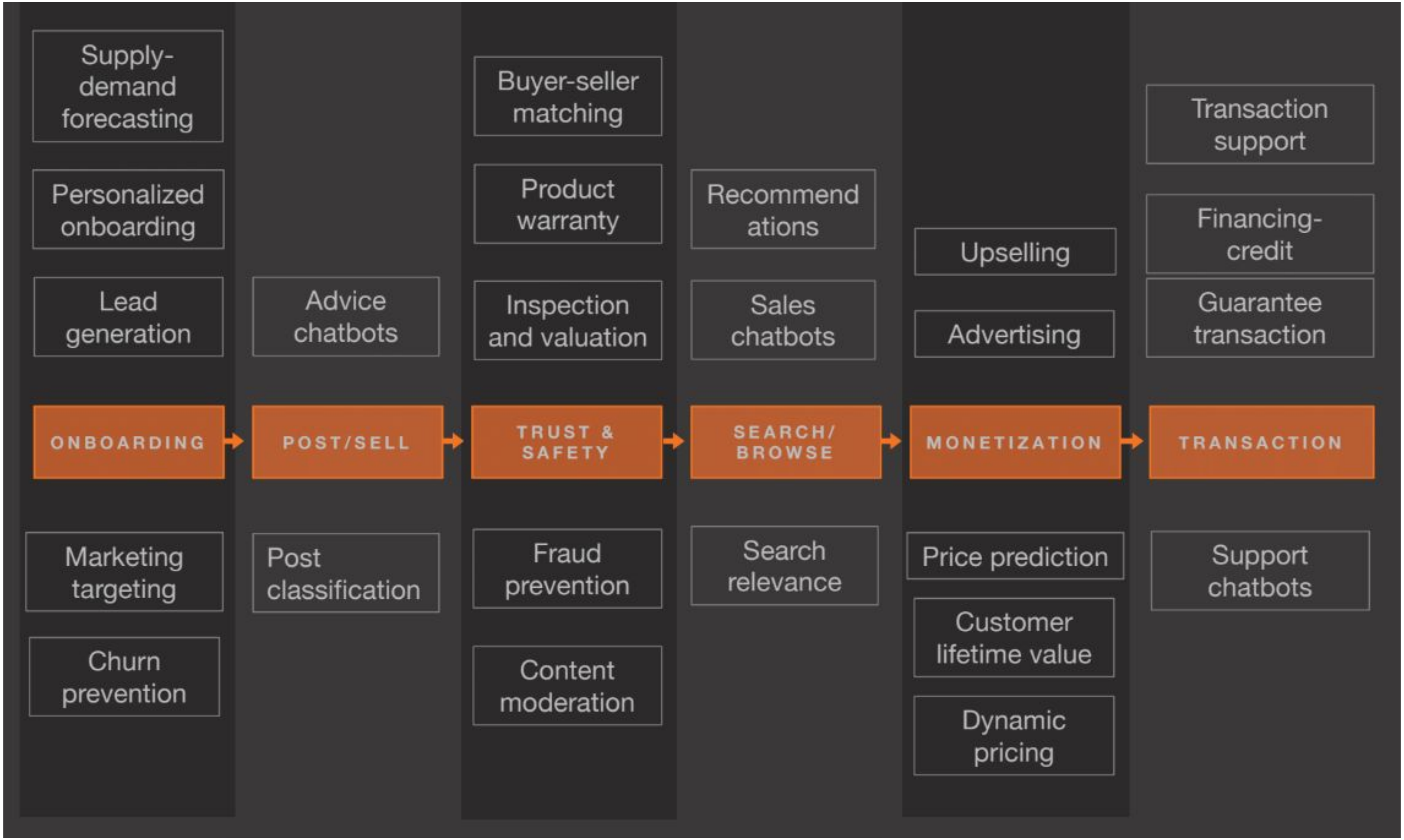


60+ million new listings per month

Market leader in 35 countries

OLX GROUP

Classifieds - Data Science areas of Application



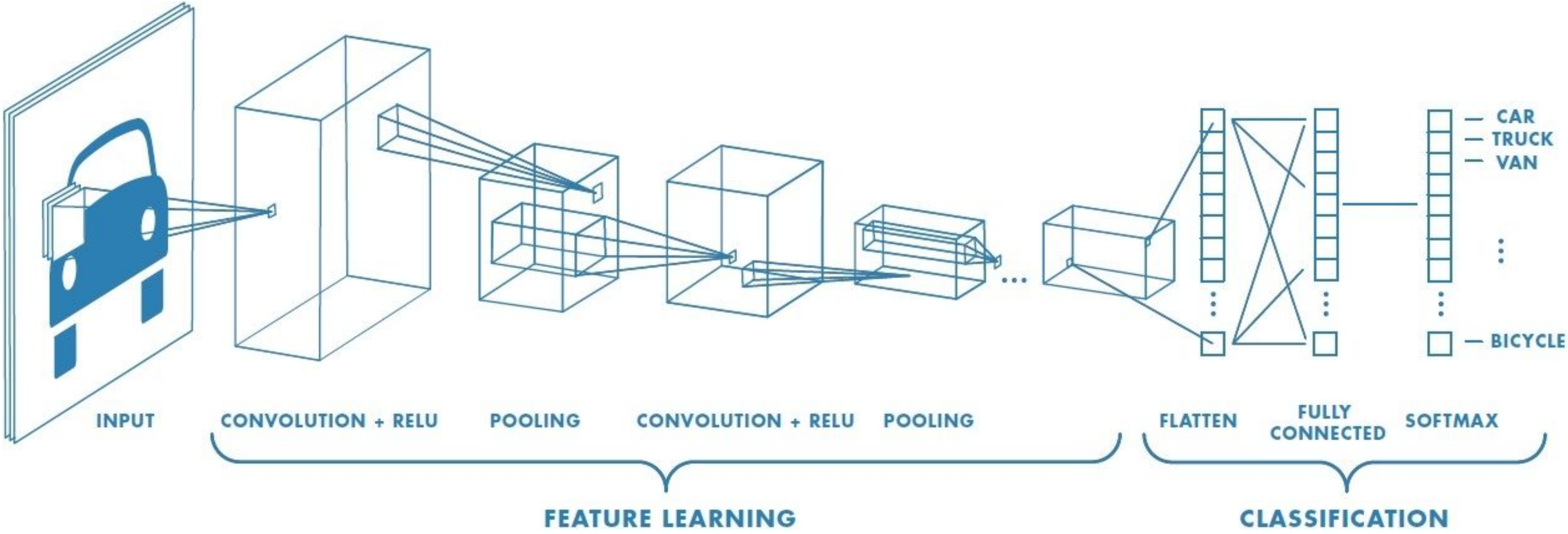
OLX GROUP

Outline

OLX GROUP

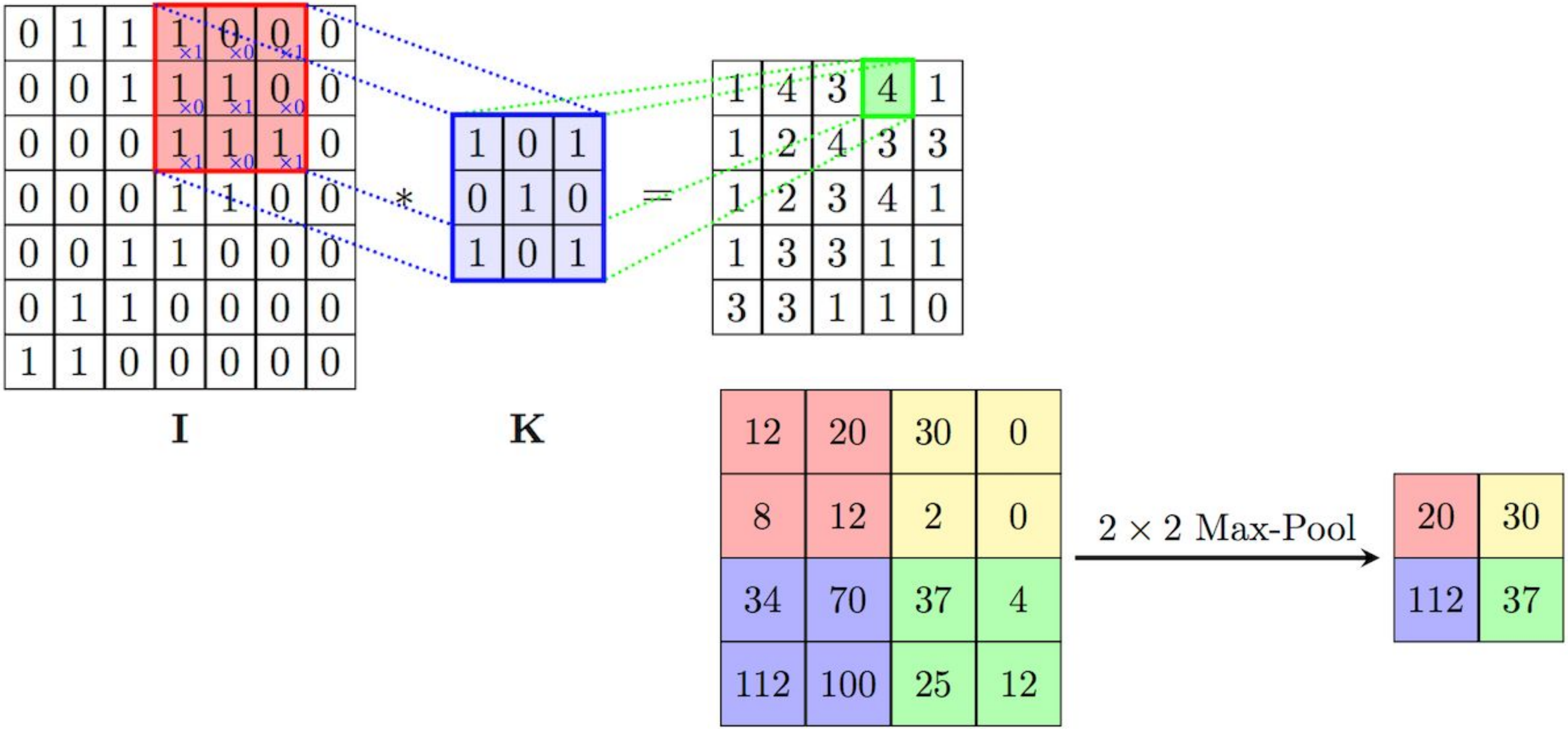
Intro to Deep Learning

Convolution Neural Networks in two images - part 1



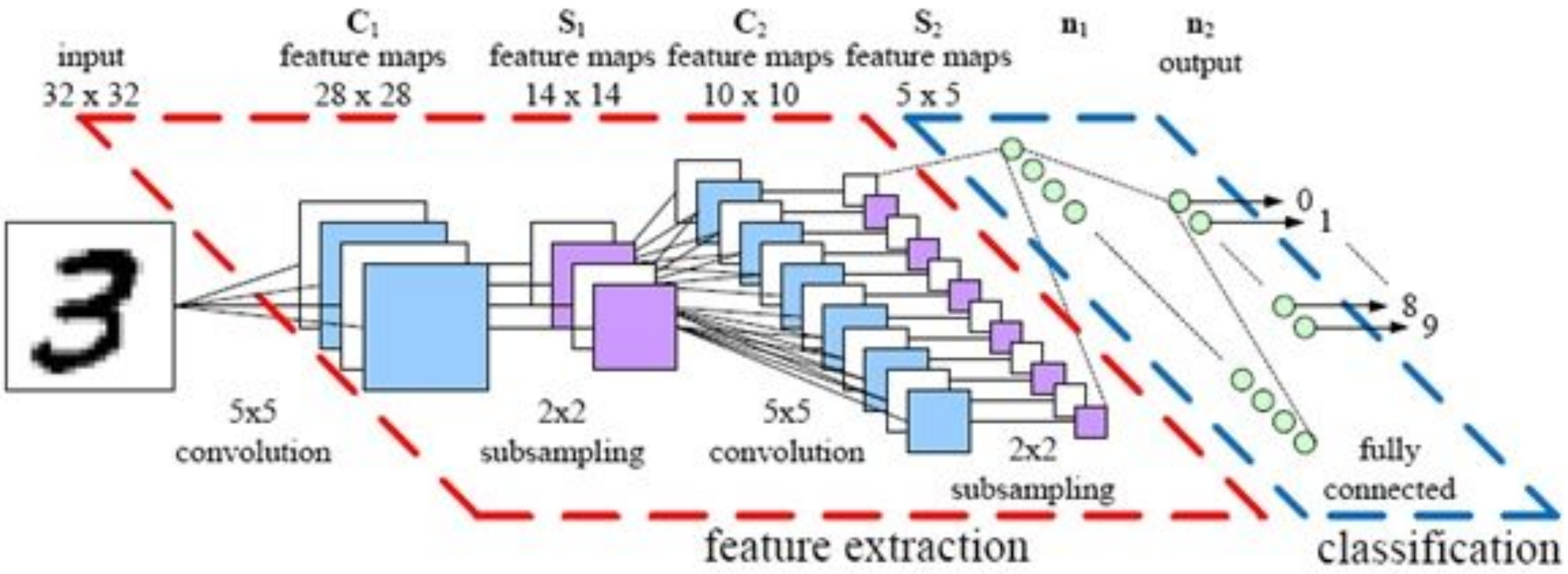
Source : https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html

Convolution and Pooling



Source : https://cambridgespark.com/content/tutorials/convolutional-neural-networks-with-keras/index.html

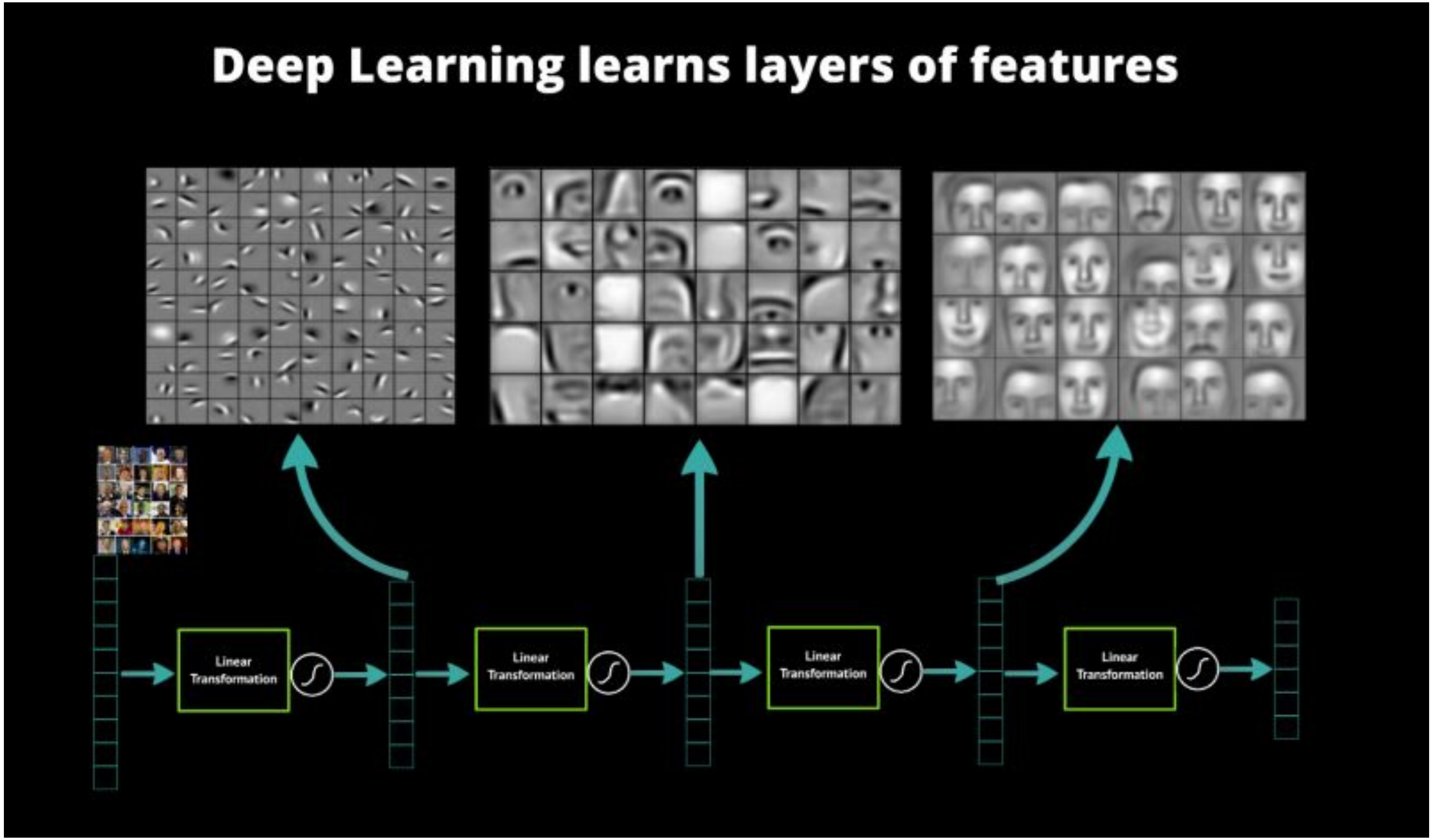
Convolution Neural Networks in two images - part 2



http://machinelearninguru.com/computer_vision/basics/convolution/convolution_layer.html

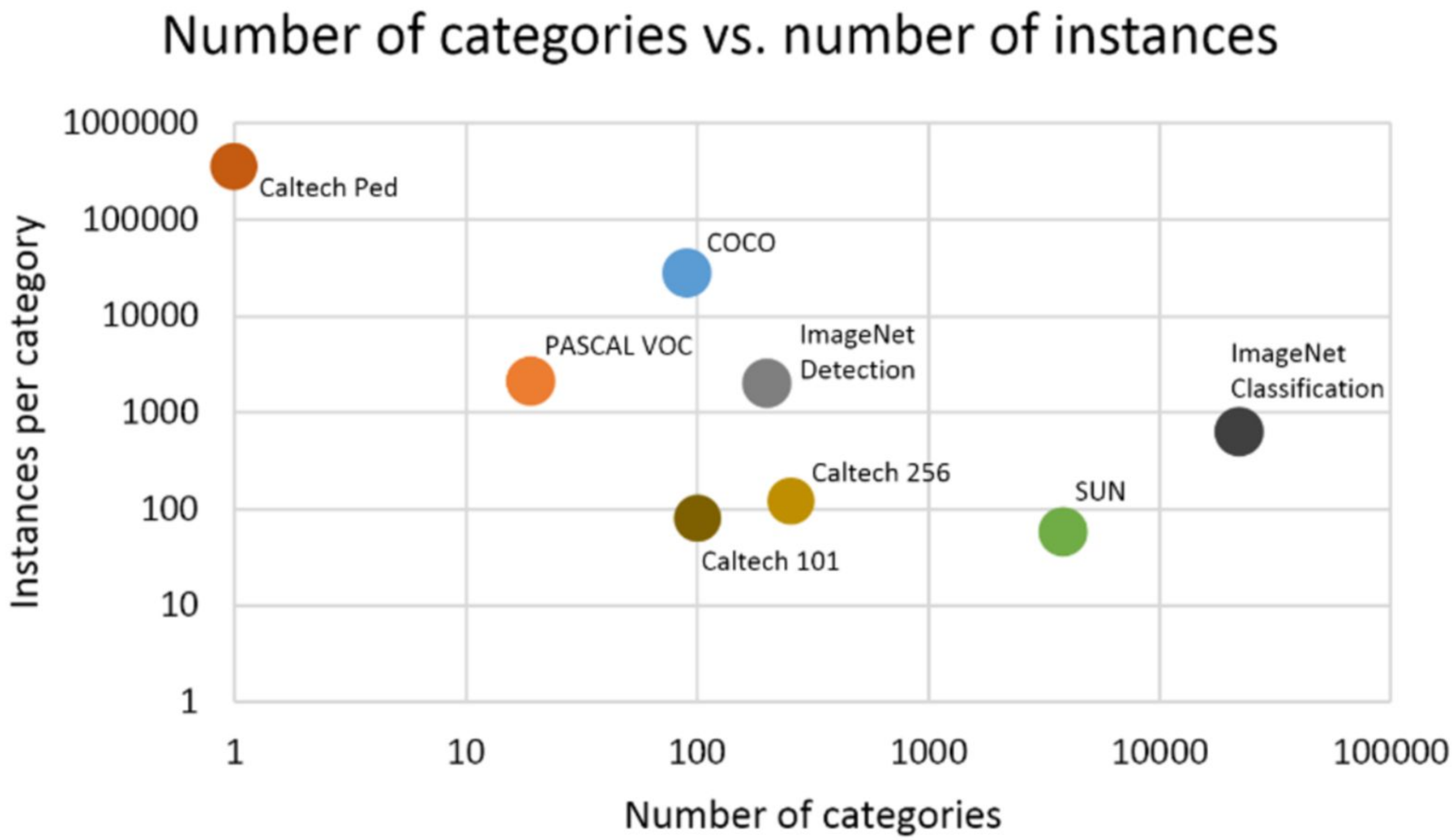
Why does it work so well?

Hierarchical building block hypothesis



Source : https://medium.com/@johnsmart/your-personal-sim-pt-4-deep-agents-understanding-natural-intelligence-7040ae074b71

Sample Sizes for Deep Learning

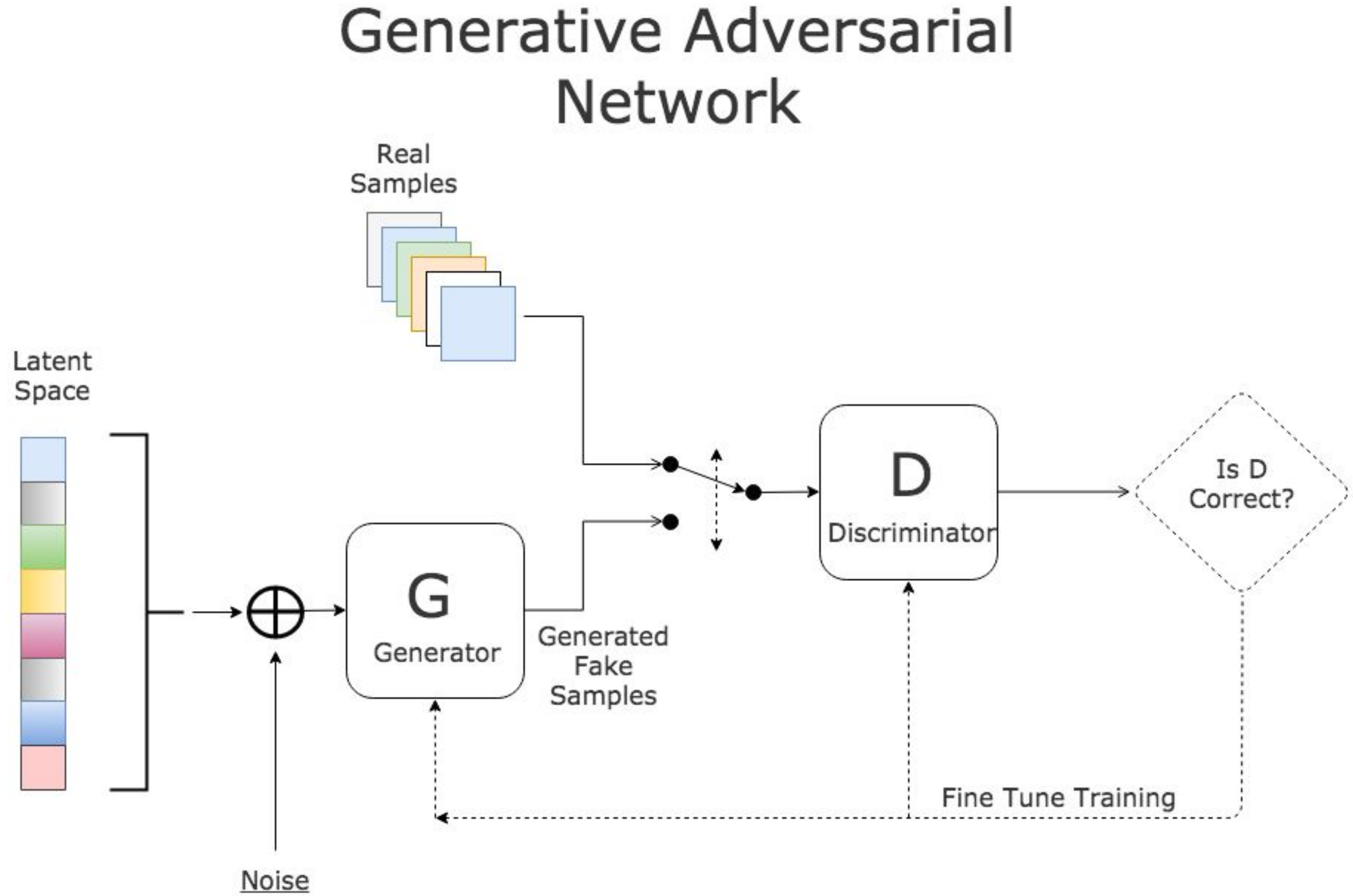


Source : https://arxiv.org/pdf/1602.07332.pdf

OLX GROUP

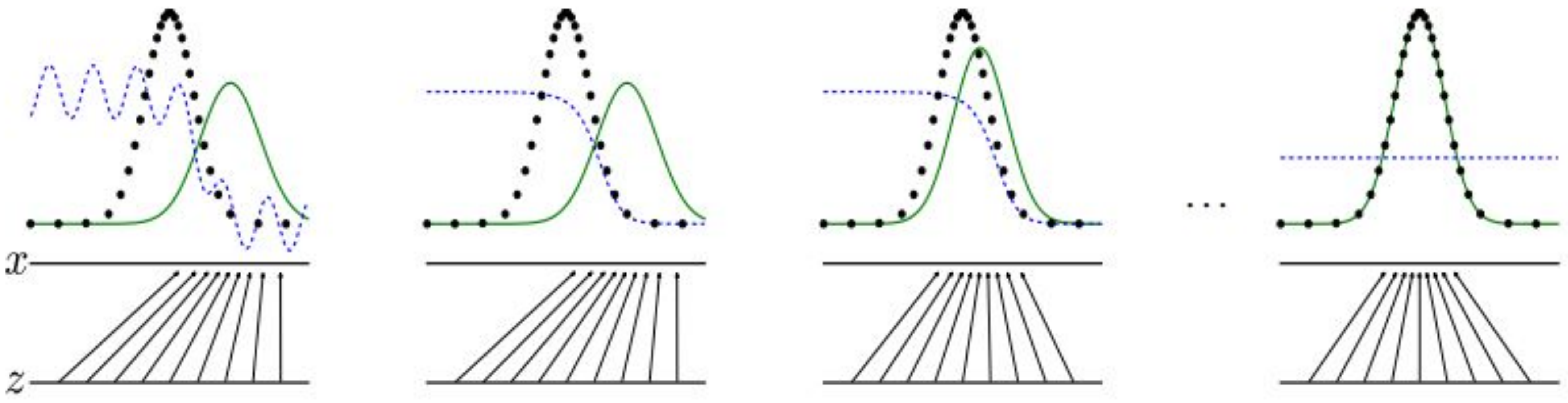
Generative Adversarial Networks

Conceptual representation of a GAN



Source : https://stats.stackexchange.com/questions/277756/some-general-questions-on-generative-adversarial-networks

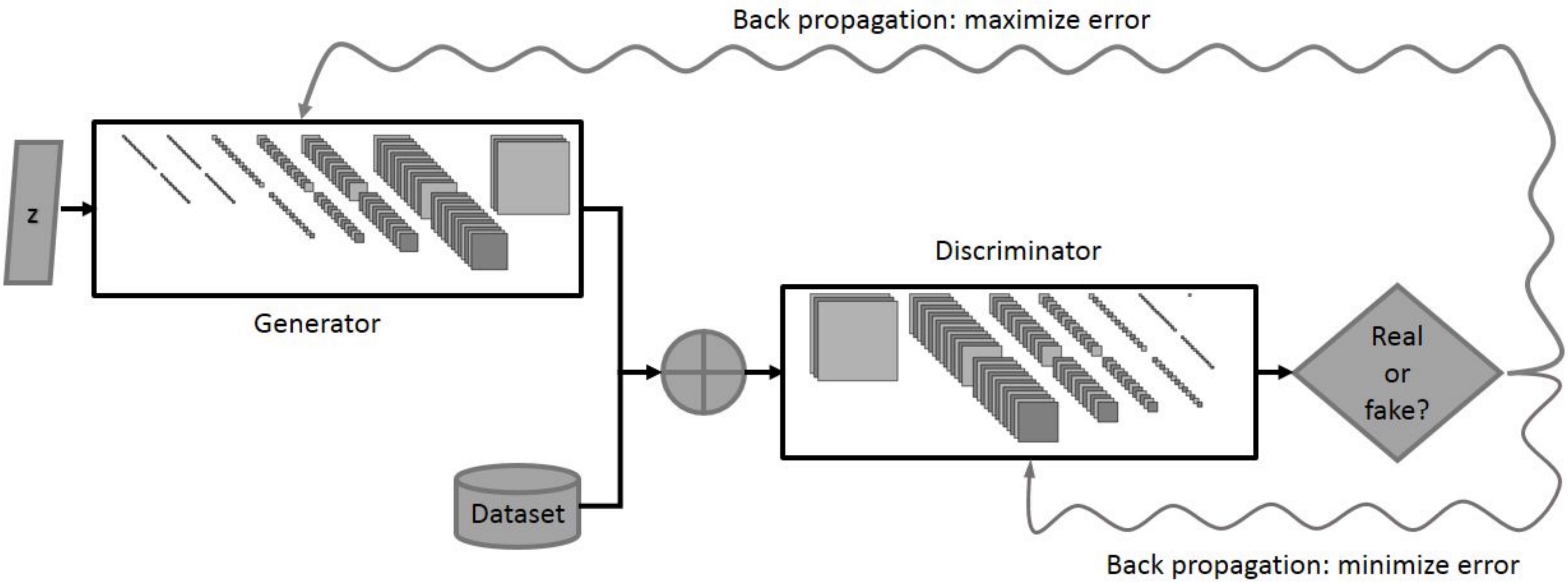
1D example of GAN



(a) (b) (c) (d)

Source : https://ireneli.eu/2016/11/17/deep-learning-13-understanding-generative-adversarial-network/

GANs in one image



Source : https://news.developer.nvidia.com/edit-photos-with-gans

Limitations of GANs (1/3)

- Mode Collapse: Generator does not generate the full range of plausible samples
- More likely if the generator is optimized while the discriminator is kept constant for many iterations

Limitations of GANs (2/3)

- GANs predict the entire sample (e.g. image) at once
- Difficult to predict pixel neighborhoods

OLX GROUP

Limitations of GANs (3/3)

- Difficult training process / convergence: Generator and discriminator keep oscillating
- Network does not converge to a stable, (near) optimal solution

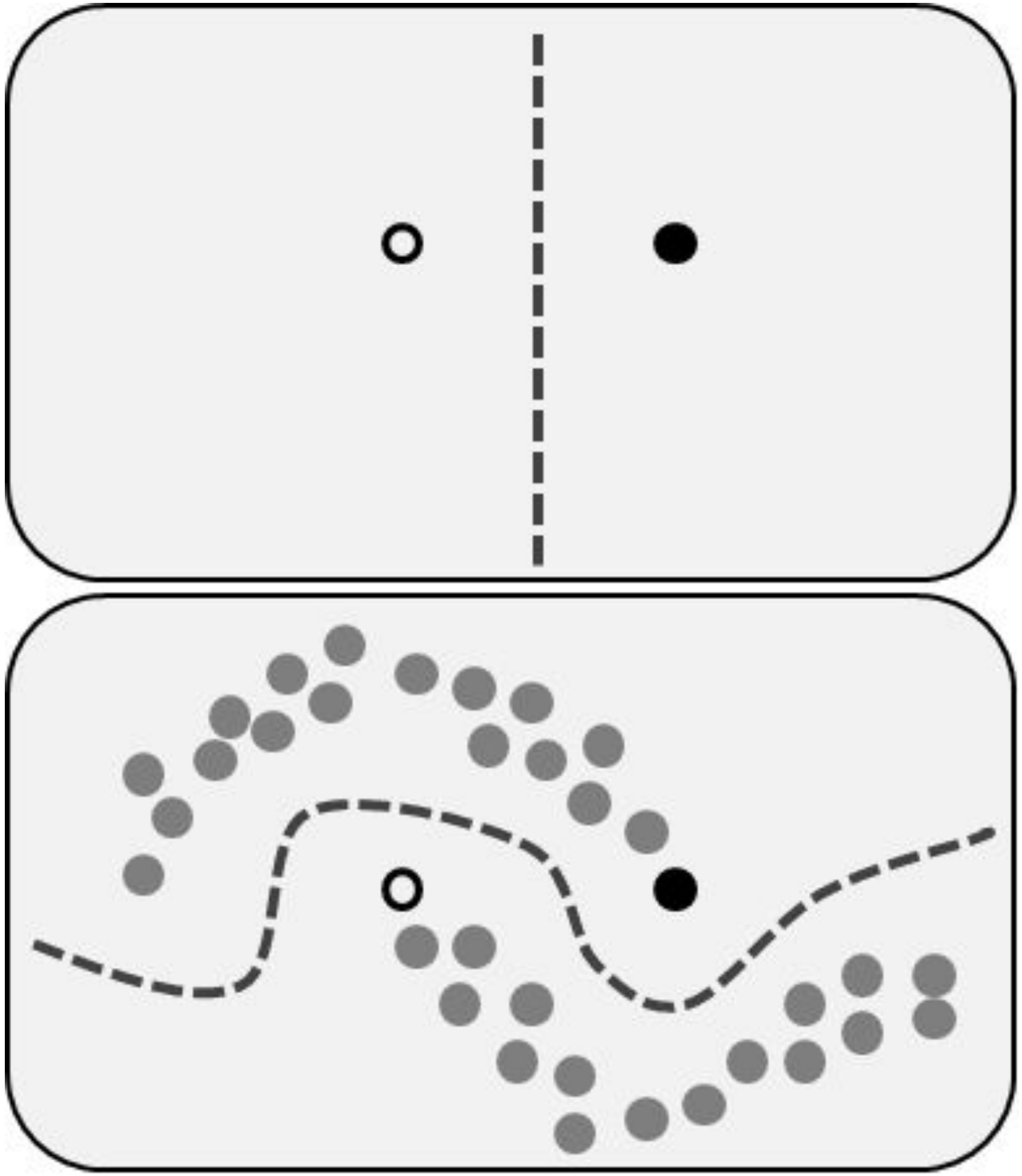
OLX GROUP

Semi-supervised learning with GANs

Why using GANs in a semi-supervised setting?

- Create a more diverse set of unlabeled data
- Get a better / more descriptive decision boundary
- Improve generalization when the amount of training samples is small

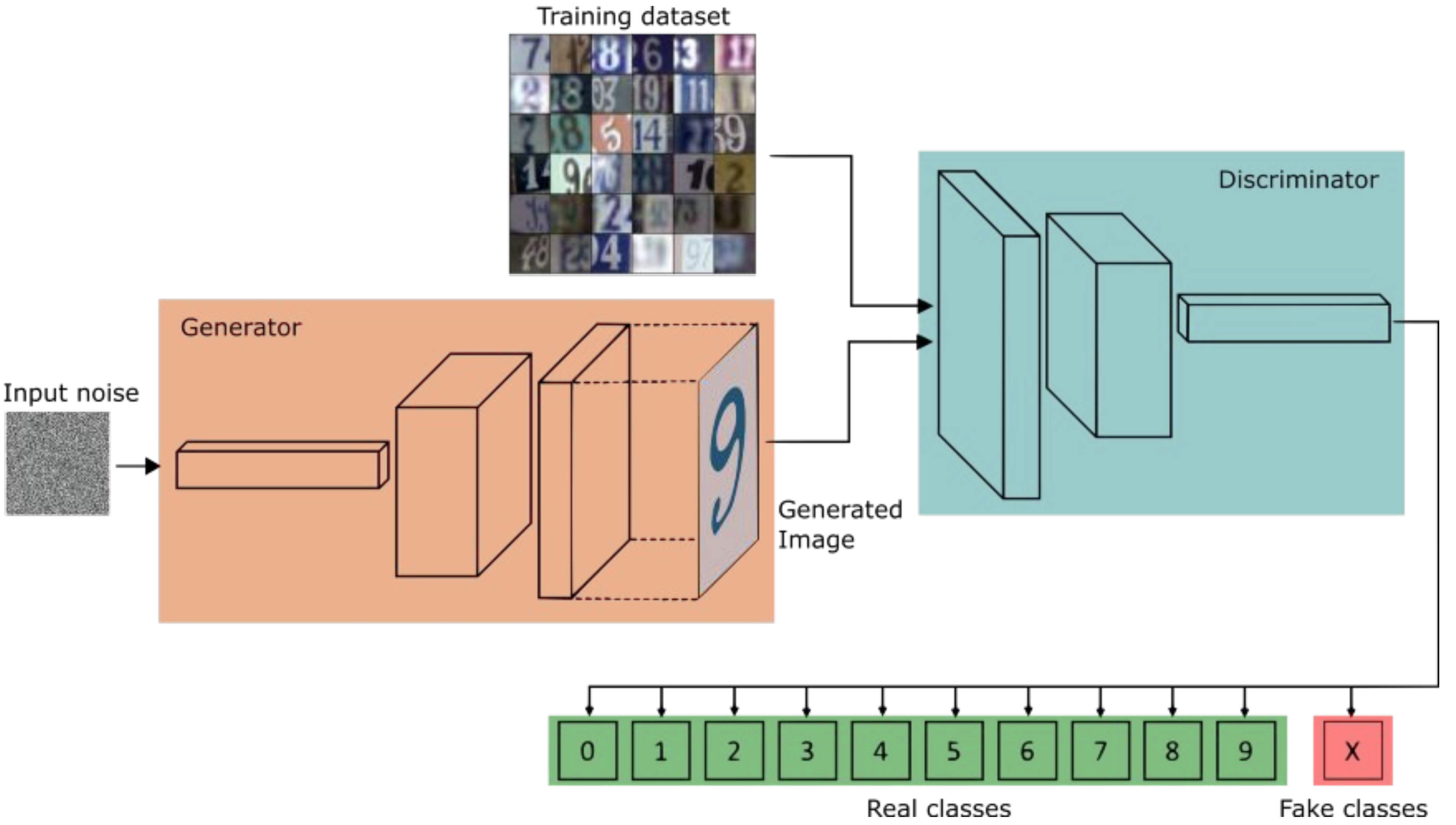
Influence of unlabeled data in semi-supervised learning



- Continuity assumption
- Cluster assumption
- Manifold assumption

OLX GROUP

Example of GAN in a semi-supervised learning role



Source : https://towardsdatascience.com/semi-supervised-learning-with-gans-9f3cb128c5e

Sources of information and their role for the discriminator

- Real samples that have labels, similar to normal supervised learning
- Real samples that do not have labels. For those, the discriminator only learns that these images are real
- Samples coming from the generator. For those, the discriminator learns to classify as fake

What do we have to modify to make it work?

- Adjust the loss functions so that we tackle both problems:
  - as per the original GAN task, use binary cross entropy for the GAN part
  - as per multiclass classification task, use softmax cross entropy (or sparse cross entropy with logits) for the supervised task
  - apply masks to ignore the label not seen in the respective subtask
  - take the average of the GAN and supervised loss

- The discriminator can use the generator's images, as well as the labeled and unlabeled training data, to classify the dataset

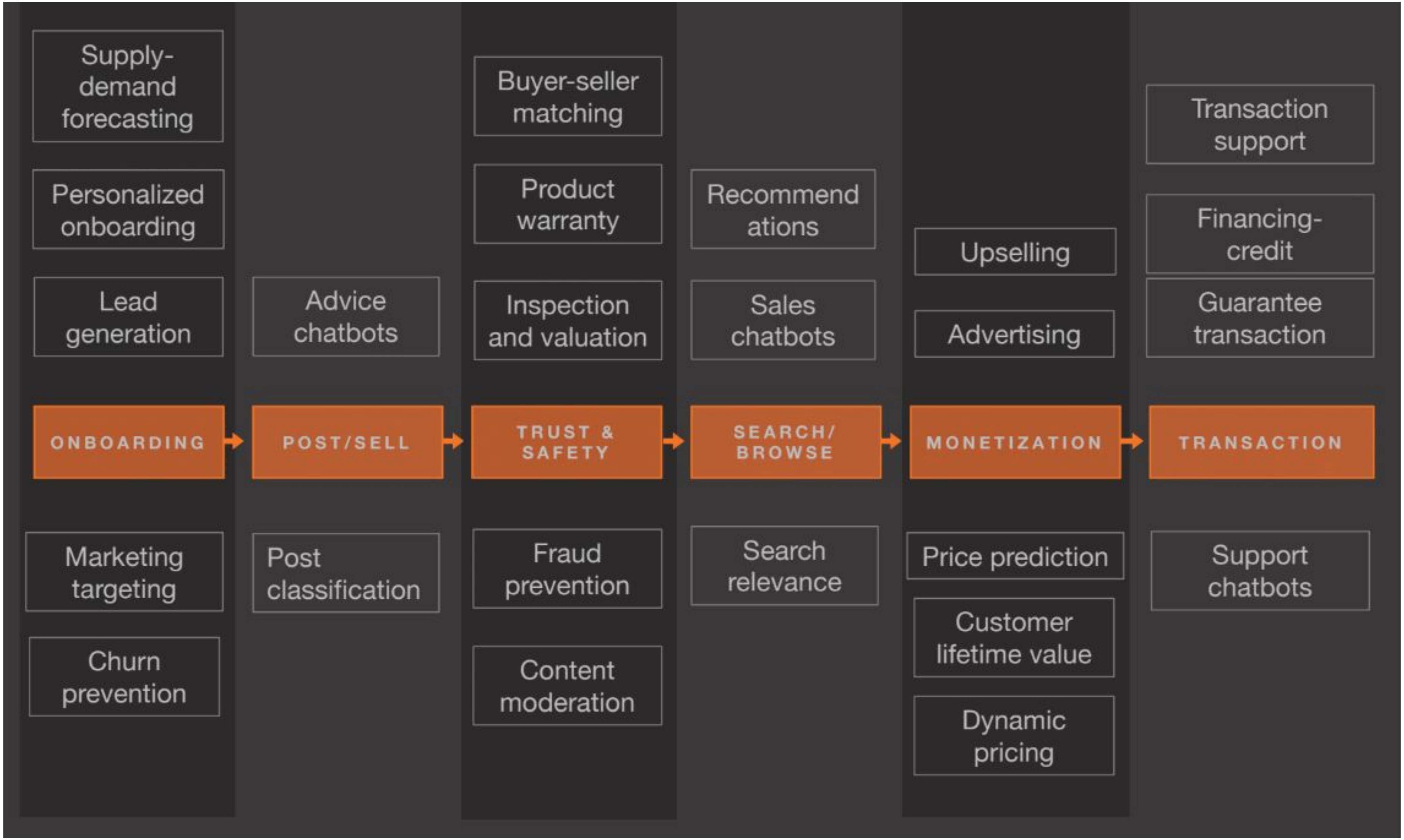
OLX GROUP

Improving the GAN part

- Training GANs is a challenging problem -> feature matching approach
    a. take the moments for some features for a “real” minibatch
    b. take the moments for same features for a “fake” minibatch
    c. minimize mean absolute error between the two

A use case from
Trust and Safety

Classifieds - Data Science areas of Application



OLX GROUP

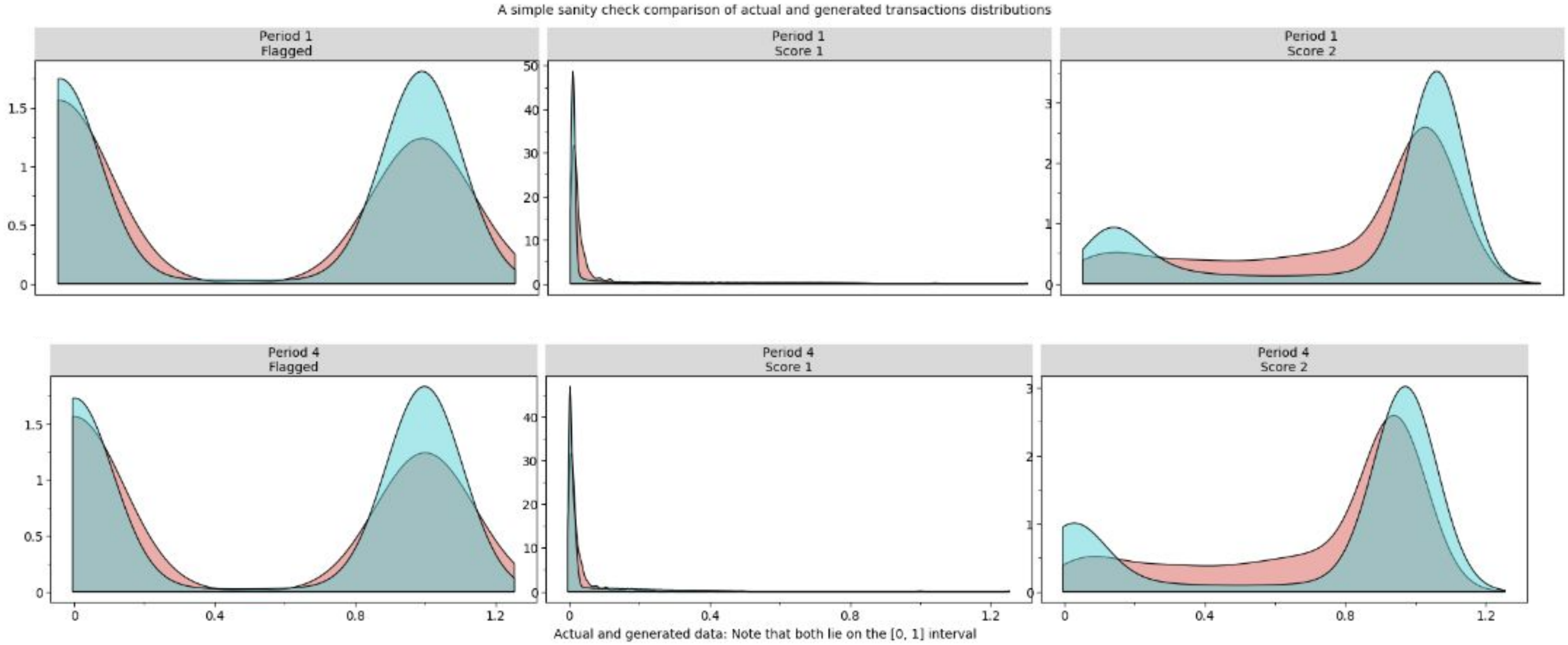
Classifieds - Data Science areas of Application

- Spam can be a problem in open platforms like classifieds
- Its annoying and can also be used as a vehicle for fraud
- How can we keep bad users out without annoying good users with stringent policies?

Why is it a challenge?

- It’s a recurring battle, with spammers becoming more inventive to circumvent the algorithms
- We may have only a few cases of cases of annotated examples for certain types of bad content

Generating Synthetic Samples with GANs



OLX GROUP

Results

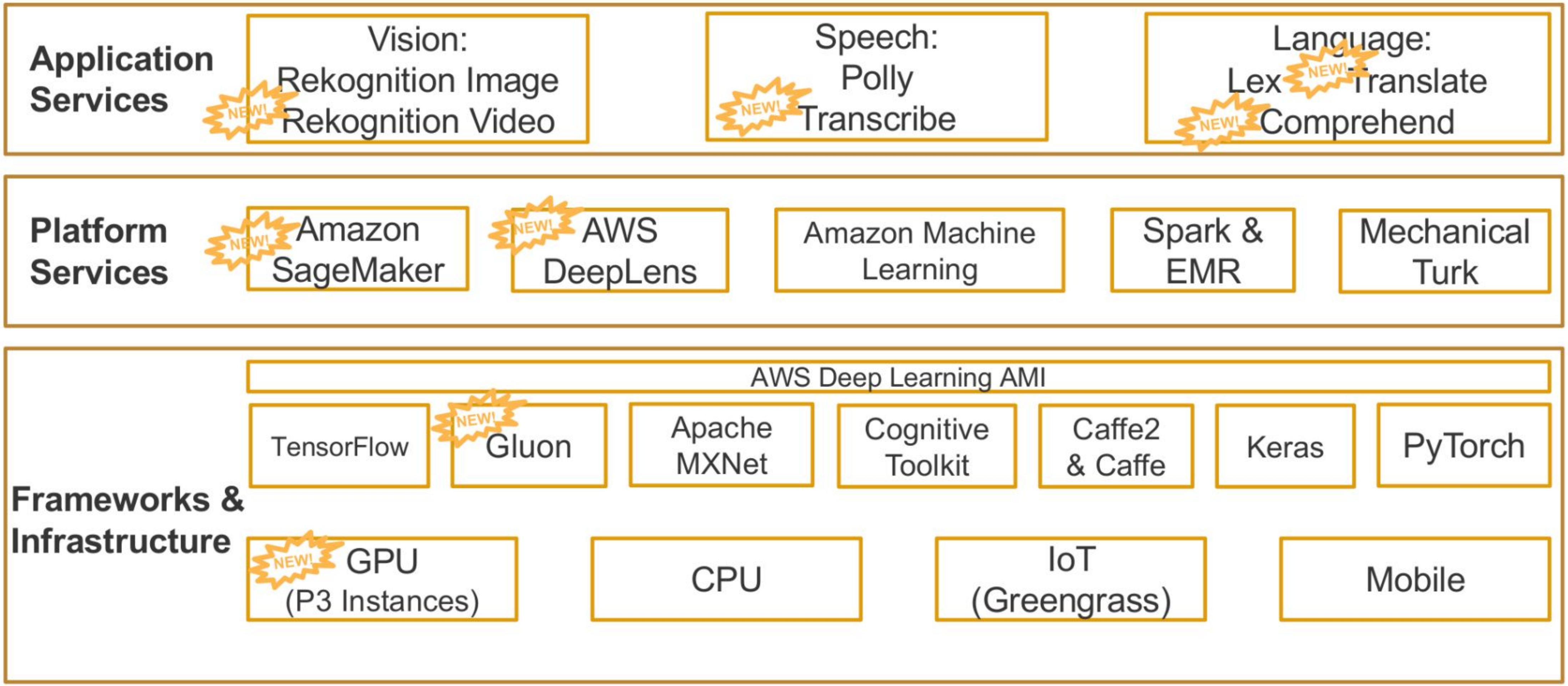
- In the presented example we only have a few tens of a certain type of spam
- We are able to quickly explore the input space and map to a latent space combining the different types of signal we have available from text, image, user behaviour, etc.
- Semi supervised learning with GANs does better (AUC = 99.9) vs Random Forest (AUC = 99.7)

OLX GROUP

Questions and Discussion

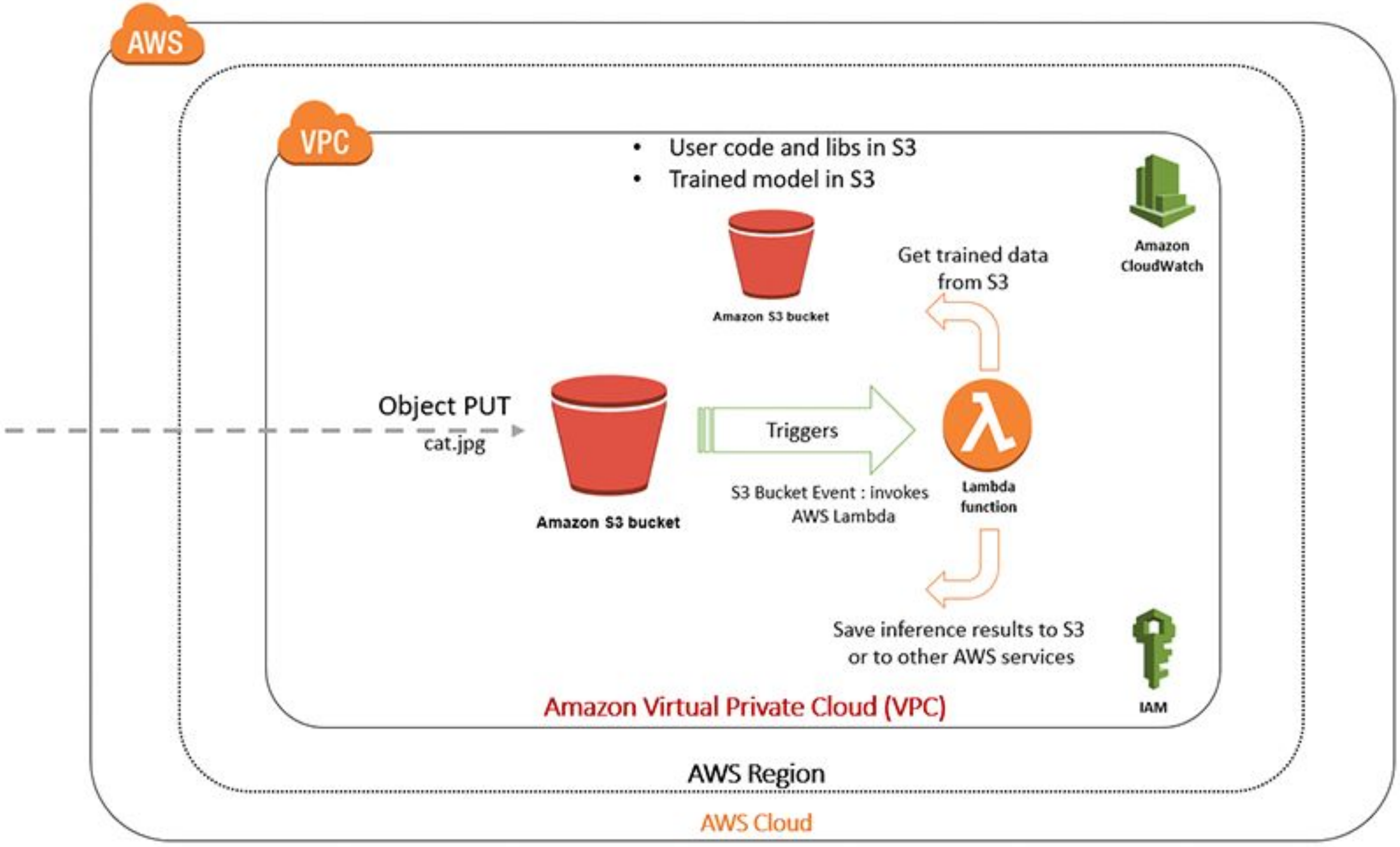
Model Deployment Options - Making endpoints

AWS Sagemaker in the ML Stack



OLX GROUP

AWS Lambda with Tensorflow



Resources

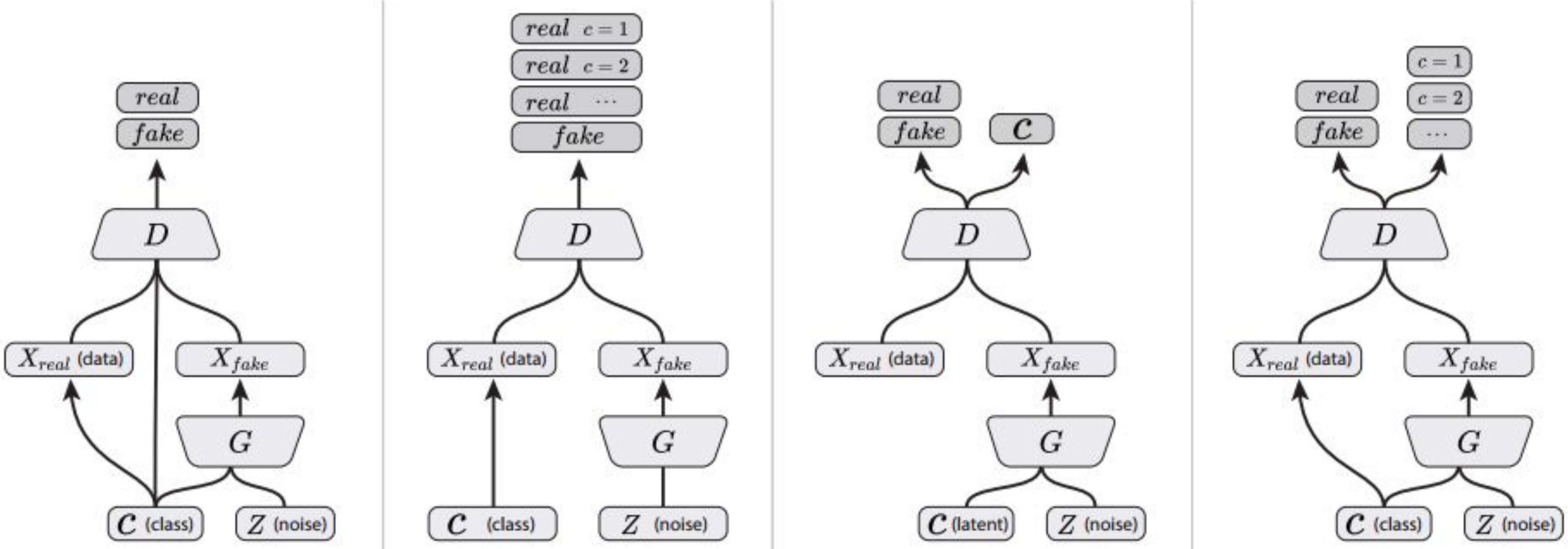
- **Using Chalice to serve SageMaker predictions**
  - https://medium.com/@julsimon/using-chalice-to-serve-sagemaker-predictions-a2015c02b033
- **SageMaker examples**
  - https://github.com/awslabs/amazon-sagemaker-examples
- **How to Deploy Deep Learning Models with AWS Lambda and Tensorflow**
  - https://aws.amazon.com/blogs/machine-learning/how-to-deploy-deep-learning-models-with-aws-lambda-and-tensorflow/

OLX GROUP

References

- **Getting started with ML (Google)**
  - https://drive.google.com/file/d/0B0phzgXZajPFZFRYNFRicjlZTVk/view
- **Semi-supervised learning with Generative Adversarial Networks (GANs)**
  - https://towardsdatascience.com/semi-supervised-learning-with-gans-9f3cb128c5e
- **CNNs, three things you need to know**
  - https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html
- **Semi-supervised learning**
  - https://en.wikipedia.org/wiki/Semi-supervised_learning
- **Variants of GANs**
  - https://www.slideshare.net/thinkingfactory/variants-of-gans-jaejun-yoo
- **From GAN to WGAN**
  - https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html

Types of GANs



Conditional GAN
(Mirza & Osindero, 2014)

Semi-Supervised GAN
(Odena, 2016; Salimans, et al., 2016)

InfoGAN
(Chen, et al., 2016)

AC-GAN
(Present Work)

Source : https://becominghuman.ai/unsupervised-image-to-image-translation-with-generative-adversarial-networks-adb3259b11b9

Deployment Options we will discuss

- **Amazon Sagemaker**
- **AWS Lambda and Tensorflow**

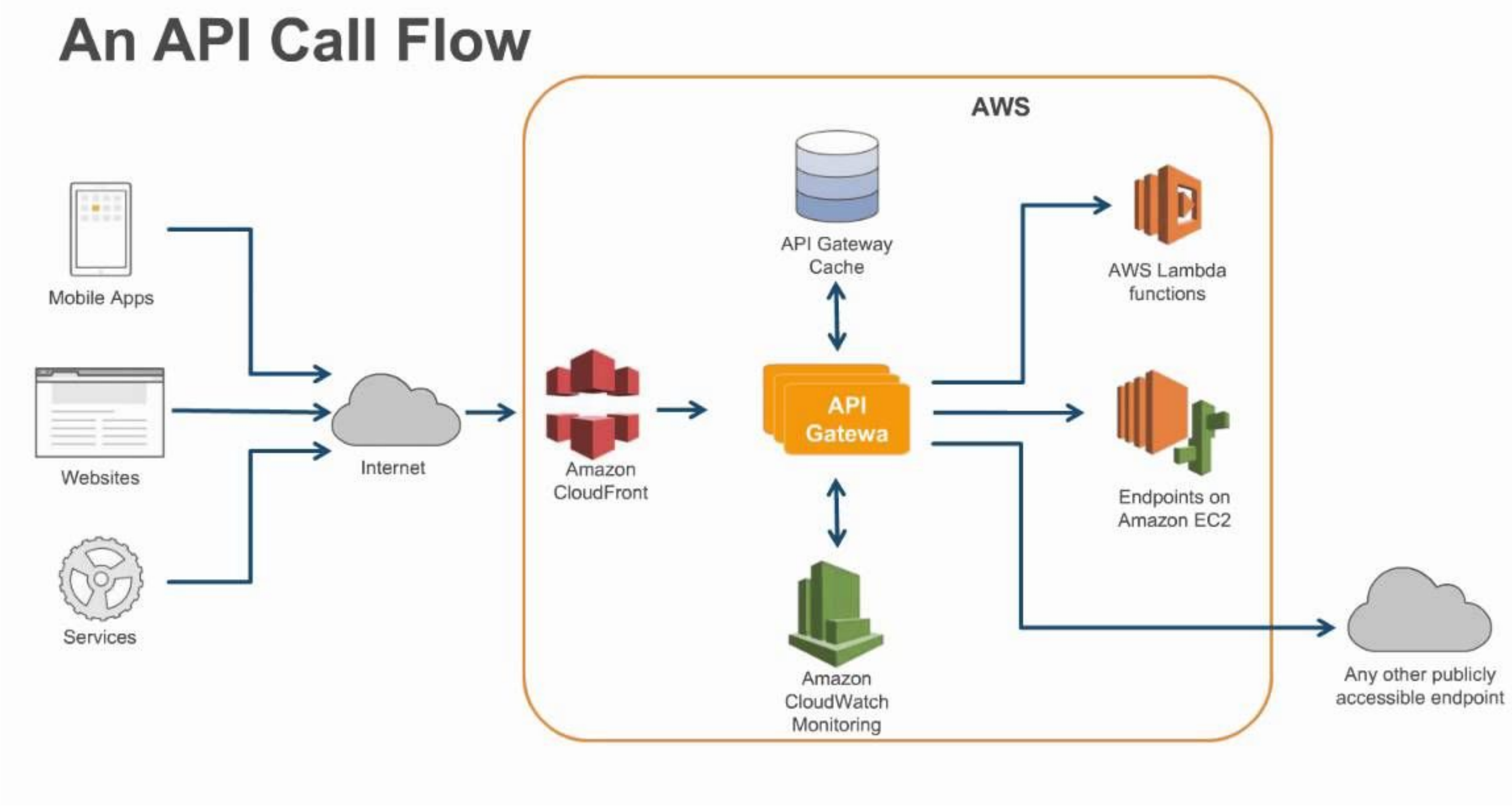
OLX GROUP

Amazon SageMaker

NEW!

A **fully managed service** that enables **data scientists** and **developers** to quickly and easily **build** machine-learning based models **into production** smart applications.

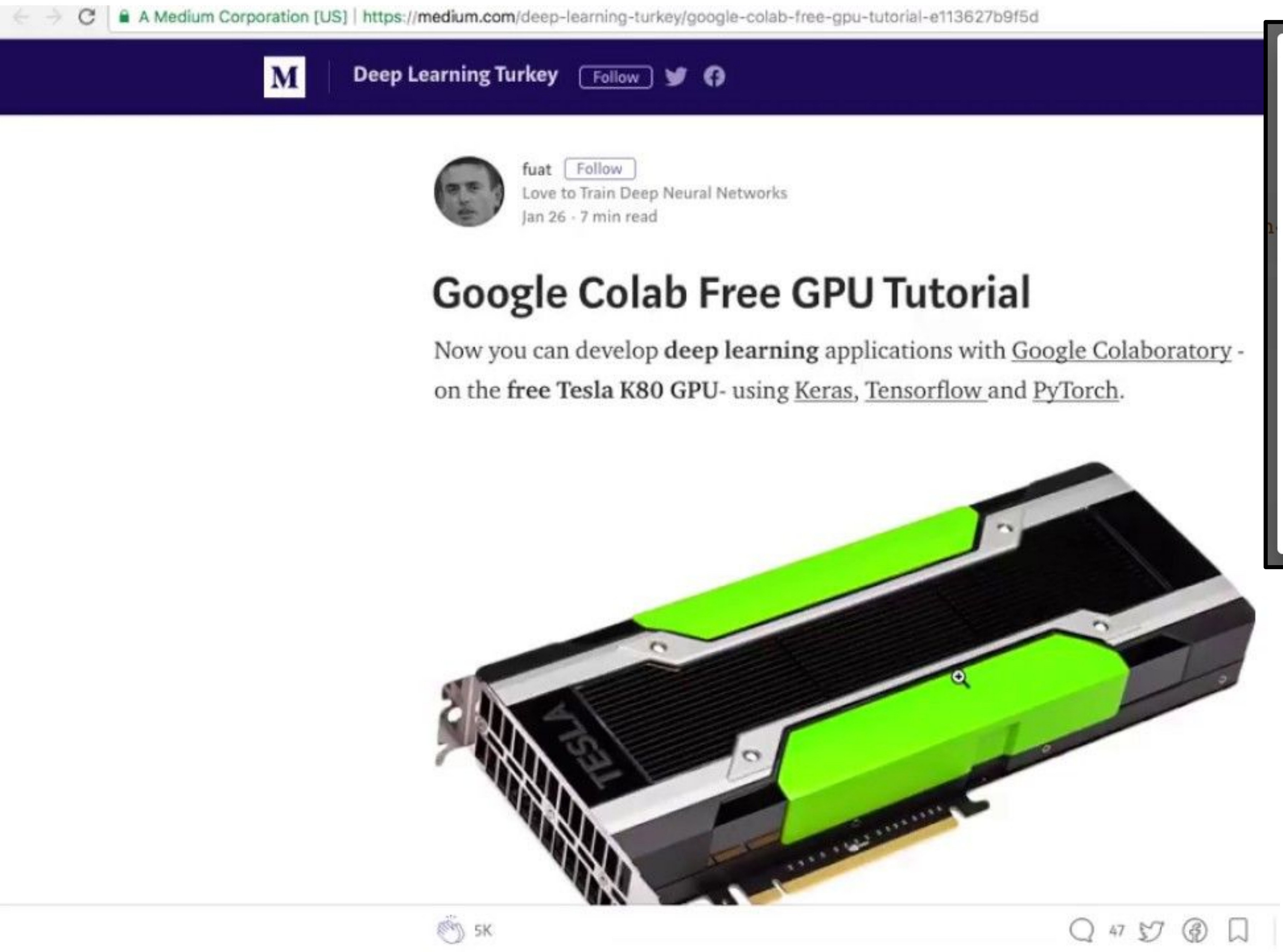
AWS Lambda

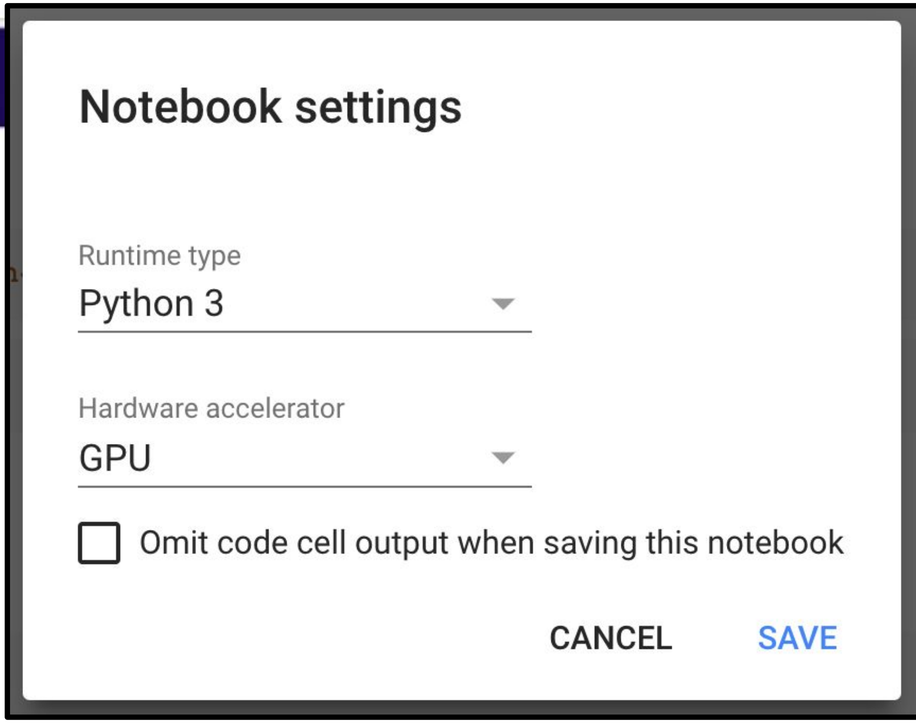


An API Call Flow
AWS
Mobile Apps
Websites
Services
Internet
Amazon
CloudFront
API Gateway
Cache
API
Gatewa
Amazon
CloudWatch
Monitoring
AWS Lambda
functions
Endpoints on
Amazon EC2
Any other publicly
accessible endpoint

Google Colaboratory

Google Colaboratory: Tensorflow + GPU + GDrive + Notebook + Free

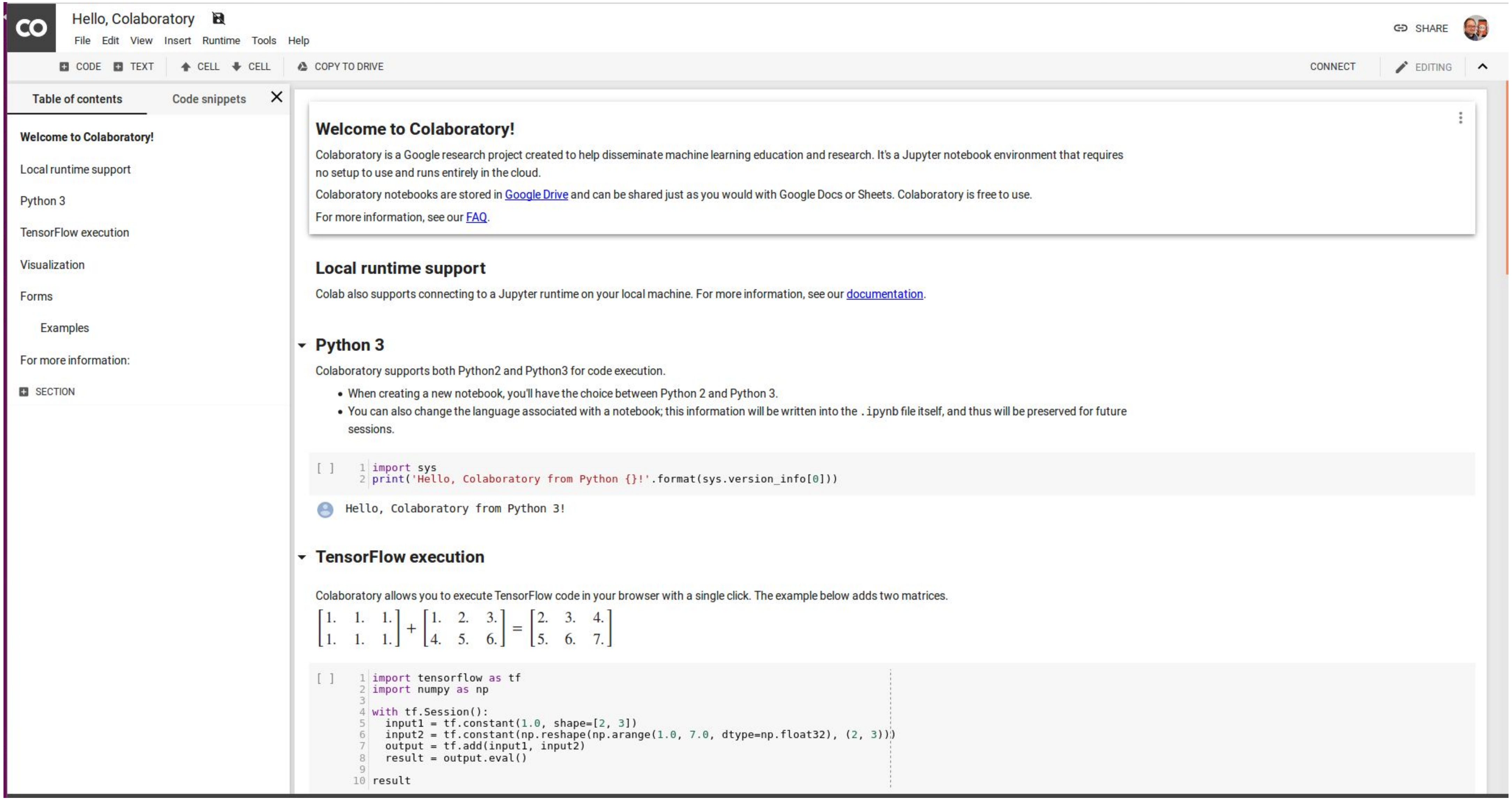




Google Colaboratory

- **What is Google Colaboratory?**
- **Tensorflow + GPU + GDrive + Notebook + Free**
- **Free (limited) use of GPUs for training**

Interface - Show me the code



https://colab.research.google.com/drive/1CqlRvhhR3bT-NUgTmgH5S9OSHn2ABVYt#scrollTo=RA4eBhnzBZO3

OLX GROUP