Машинное обучение в прикладных задачах, решаемых в Лаборатории информационных технологий им. М.Г. Мещерякова

Александр Ужинский
auzhinskiy@jinr.ru

17.11.2022 Дубна

1. Распознавание болезней и проблем в развитии растений

2. Контроль загрязнения тяжелыми металлами

Industry 4.0

The term "Industry 4.0" is used to signify the beginning of the fourth industrial revolution – the previous three being mechanical production, mass production, and then the digital revolution. It could be argued that Industry 4.0 is simply an amalgamation of the three previous eras in manufacturing, but Industry 4.0 is poised to be much more impactful than that.

Industrial Internet of Things (IIoT), Automation, Artificial Intelligence, Big Data & Analytics, The Cloud, Cybersecurity, Simulations, Robotics, Smart manufacture, Mobile devices, Smart manufacture, etc.

Распознавание болезней и проблем с развитием растений

Introduction

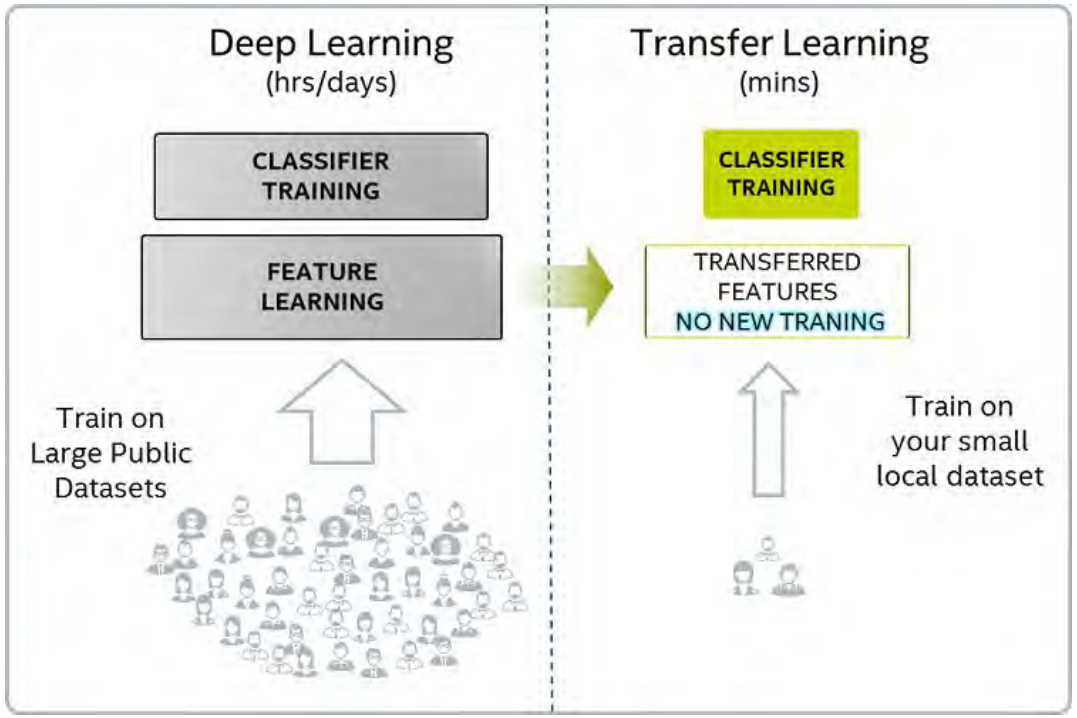
Crop losses are a major threat to the wellbeing of rural families, to the economy and governments, and to food security worldwide
USAblight (a national project on late blight on potato and tomato) says that (annual) global losses '**exceed US$6.7 billion**'.

Globally, about **16% of all crops are lost** to plant diseases each year.
Dr. Caitilyn Allen Department of Plant Pathology, University of Wisconsin–Madison

Increasing number of smartphones and advances in deep learning field opens **new opportunities** in the crop diseases detection.

The aim of our research is to **facilitate the detection** and preventing diseases of agricultural plants by both **deep learning and programming services**. The idea is to develop multifunctional platform that will use modern organization and deep learning technologies to provide new level of service to farmer's community.

PDDP Why do we need the database?



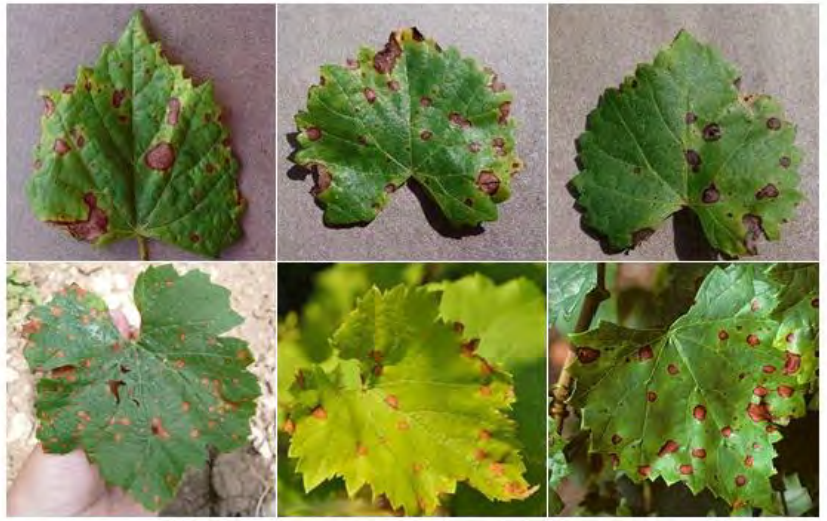
Steps of general approach:
1. take a deep network pretrained on a big dataset;
2. fine-tune the chosen deep classifier on the huge images-database (PlantVillage);
3. evaluate it on a test subset of images, collected from the Internet.
ResNet50 architecture showed the best result:

- accuracy on a test subset of the PlantVillage data – 99.4%;
- accuracy on 30 images collected from the Internet – 48%.

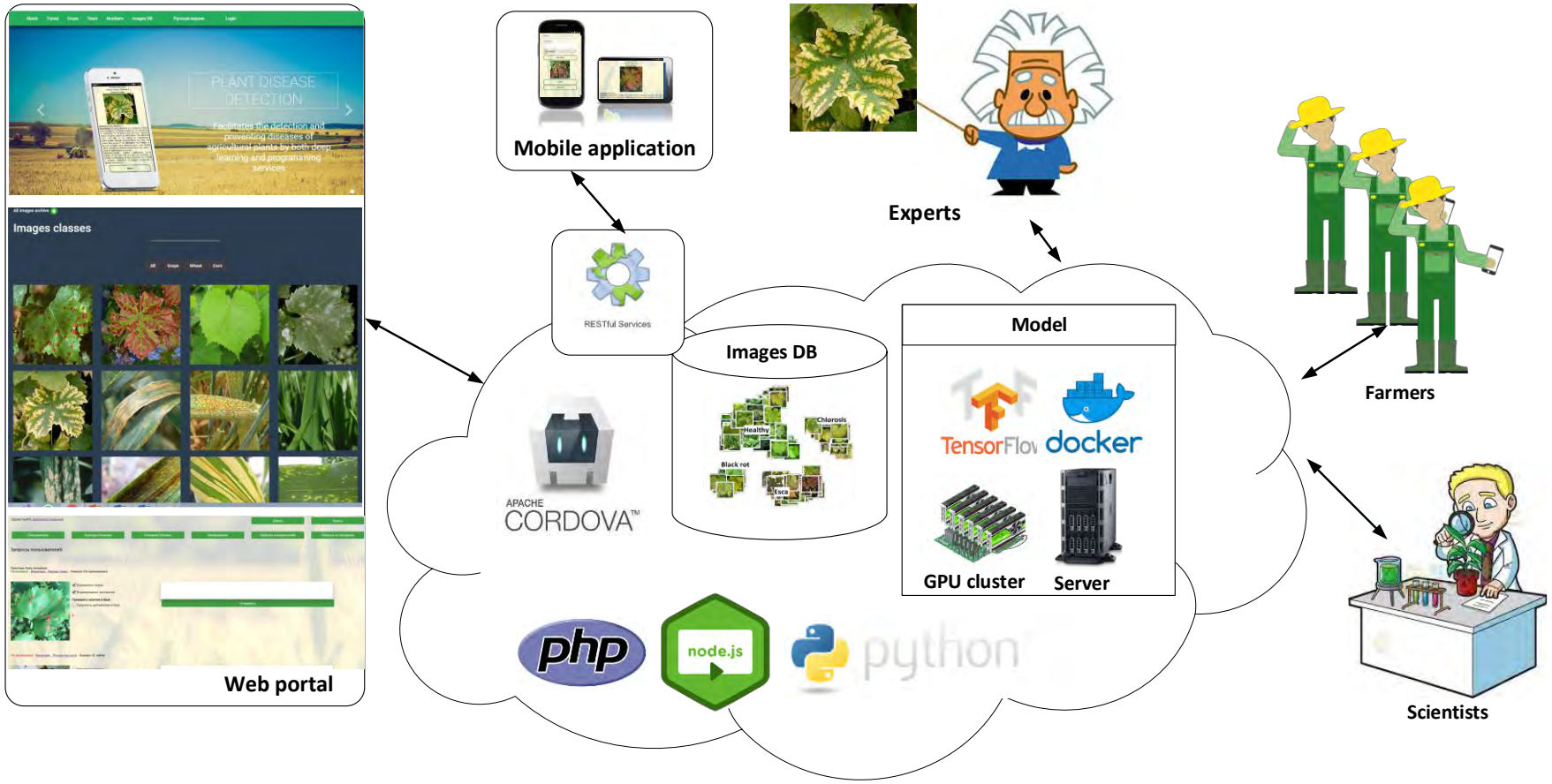
Transfer learning was not effective, but why?

Attentively look on this picture. First row – PlantVillage images, second row – real-life images from the Internet.

Do you see anything strange?

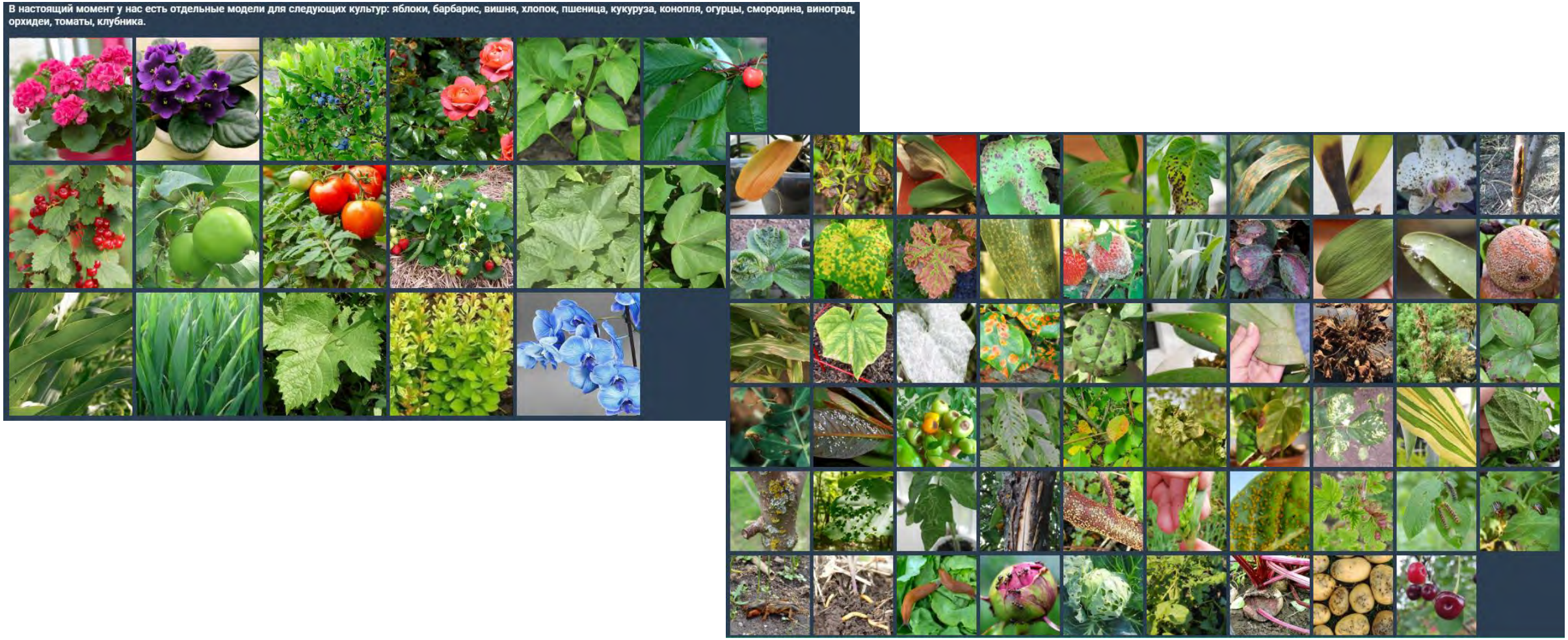


PDDP Architecture



PDDP consists of a set of interconnected services and tools developed, deployed and hosted with the help of the JINR cloud infrastructure. Our web-portal (pdd.jinr.ru), was developed with the Node.js and PHP. It provides not only a web-interface but also the API for third-party services. We have the TensorFlow model in the Docker realized as a Tensorflow serving. The model can work at the virtual server, or at a GPU cluster.
We have a mobile App for Android that was developed using the Flutter, so we could build it for iOS, and Windows.

PDDP database (http://pdd.jinr.ru)

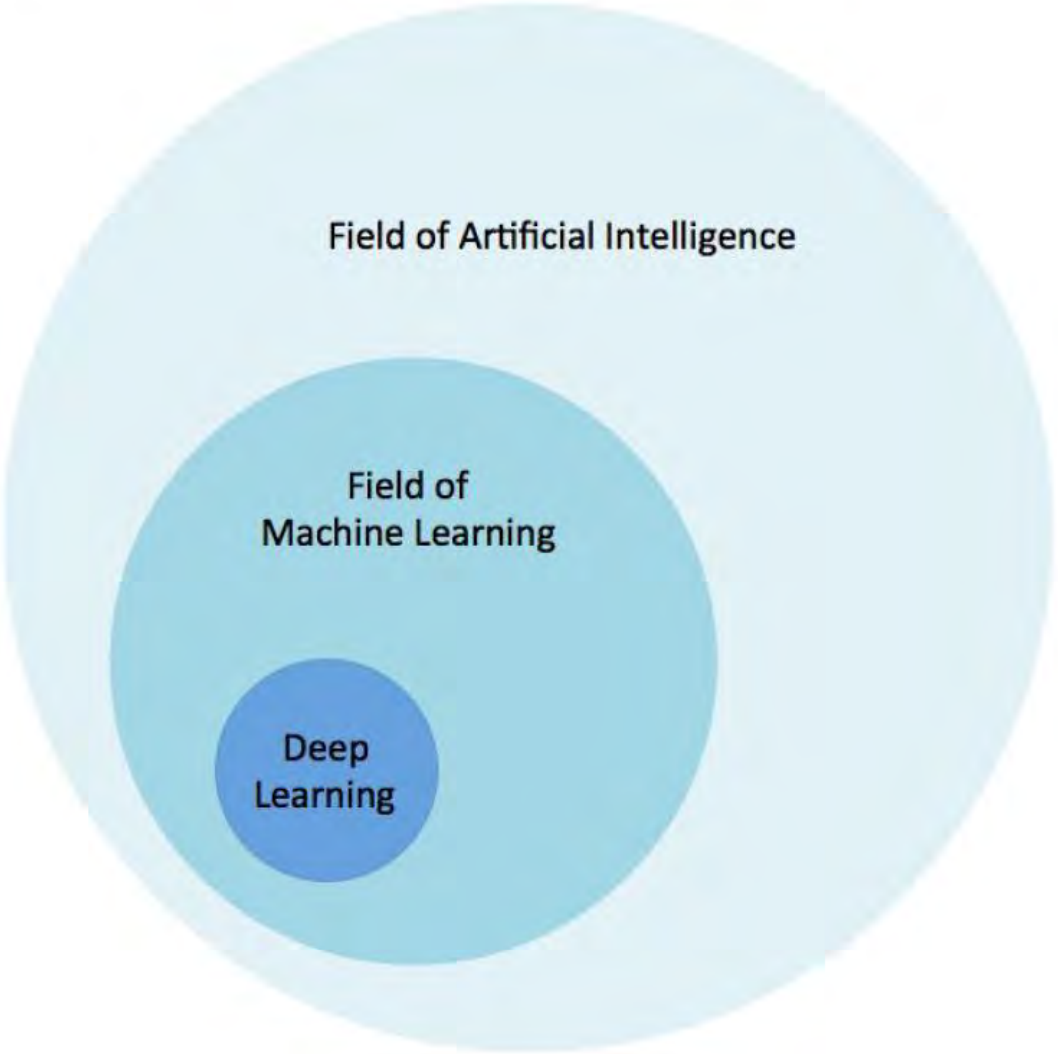


Частные модели: яблоки, барбарис, вишня, хлопок, пшеница, кукуруза, огурцы, смородина, виноград, орхидеи, томаты, клубника.

Болезни – более 40 видов

Basis

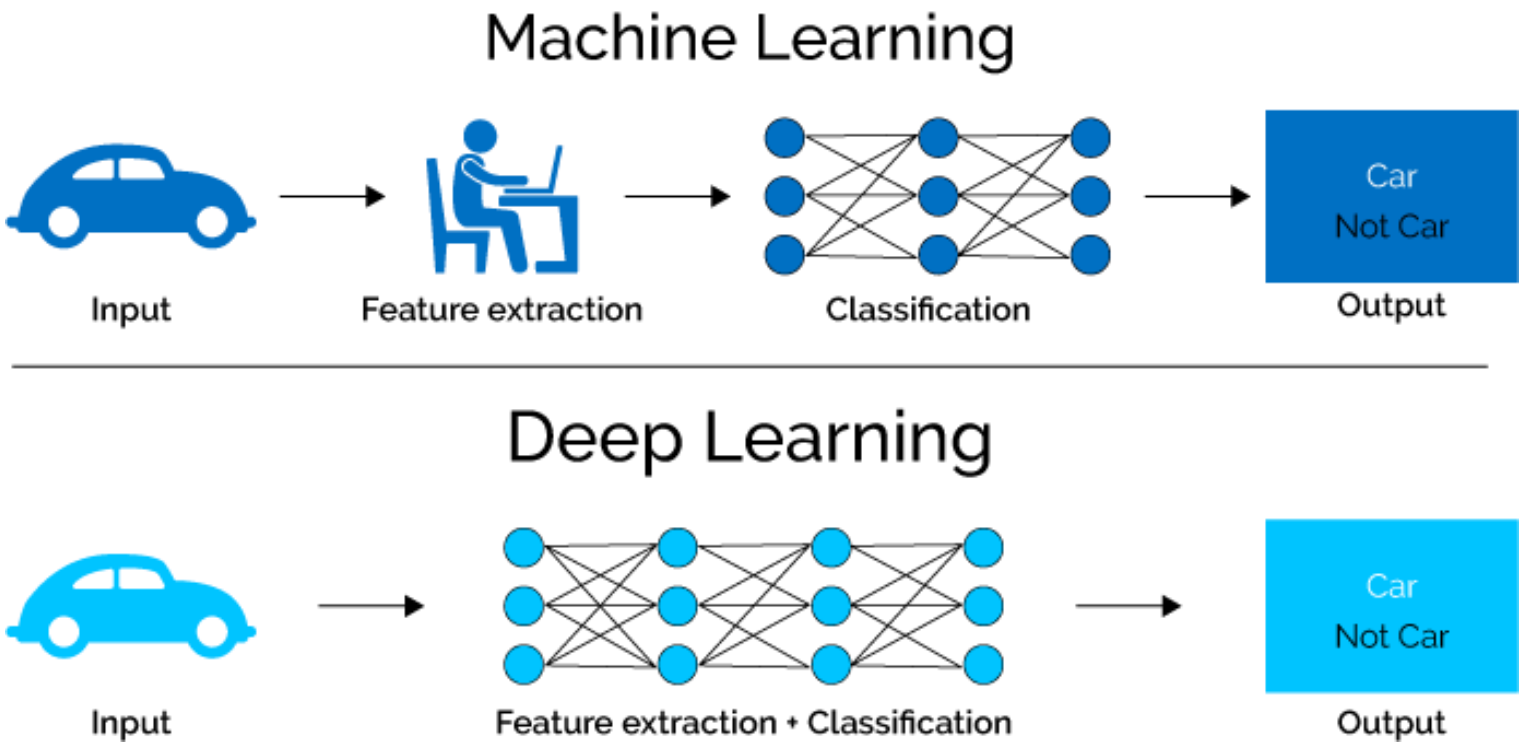
Место Deep Learning среди других областей



ML vs. Deep Learning

Deep learning (DL) is a machine learning subfield that uses multiple layers for learning data representations

DL is exceptionally effective at learning patterns



Machine Learning Types

Supervised: learning with labeled data

Example: email classification, image classification

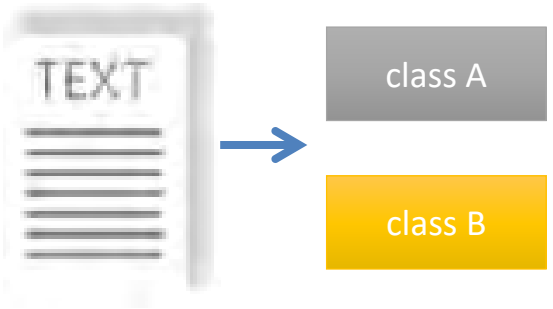
Example: regression for predicting real-valued outputs

Unsupervised: discover patterns in unlabeled data

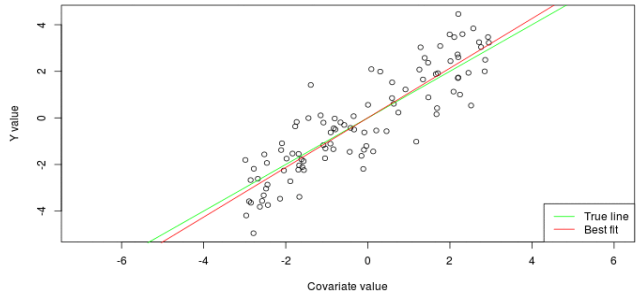
Example: cluster similar data points

Reinforcement learning: learn to act based on feedback/reward

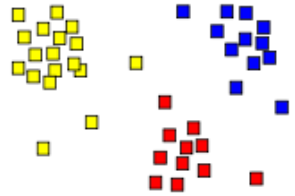
Example: learn to play Go



Classification



Regression

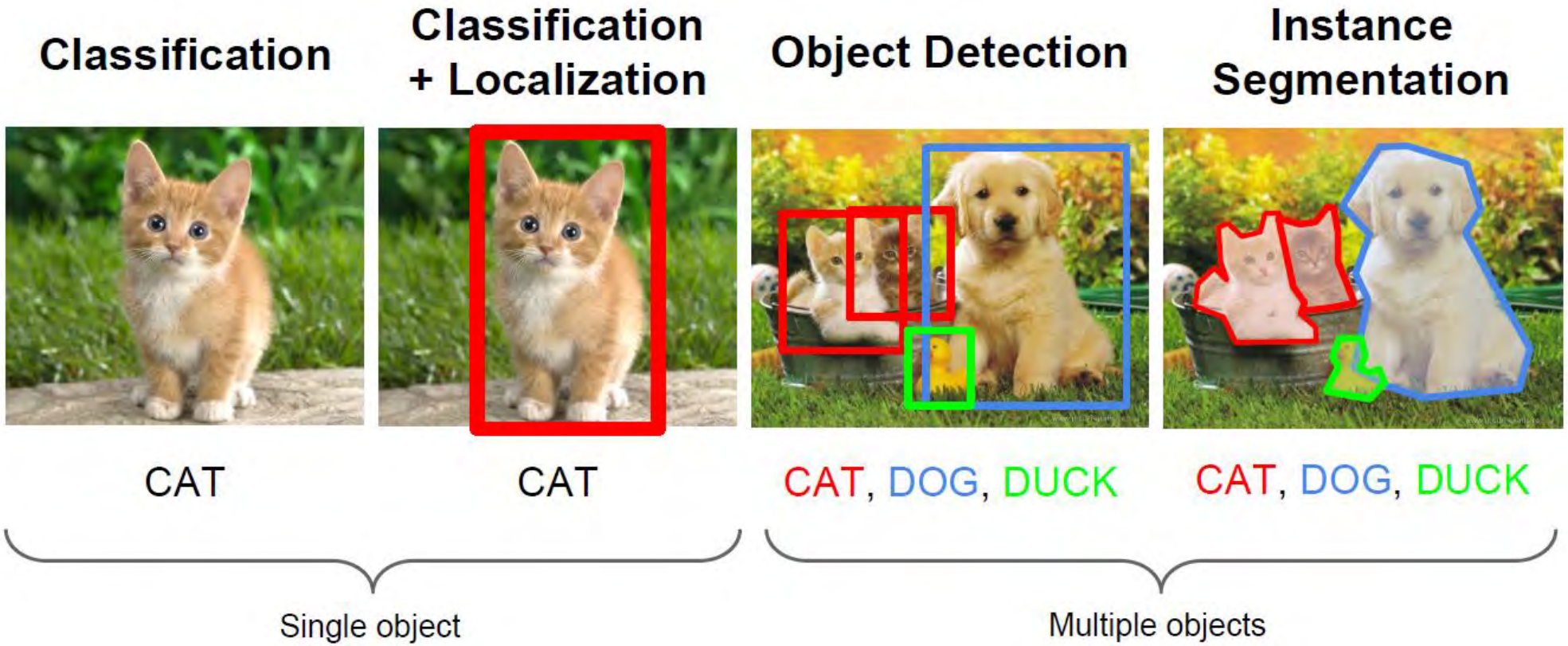


Clustering

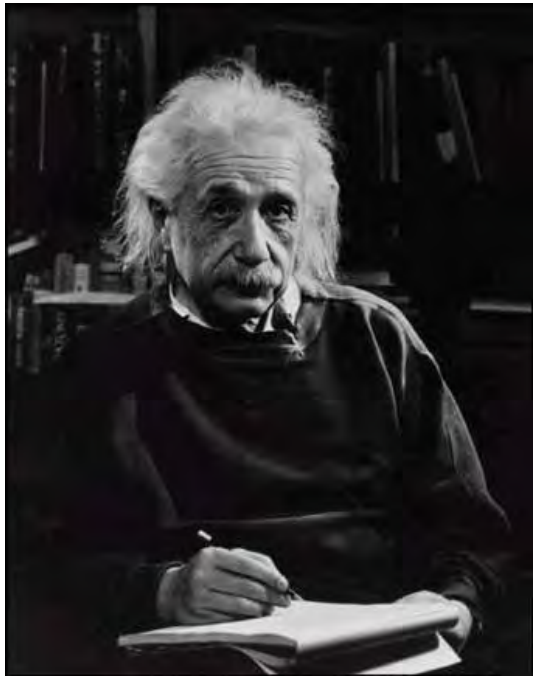
Computer Vision Tasks

Computer vision has been the primary area of interest for ML

The tasks include: classification, localization, object detection, instance segmentation



Computer Vision

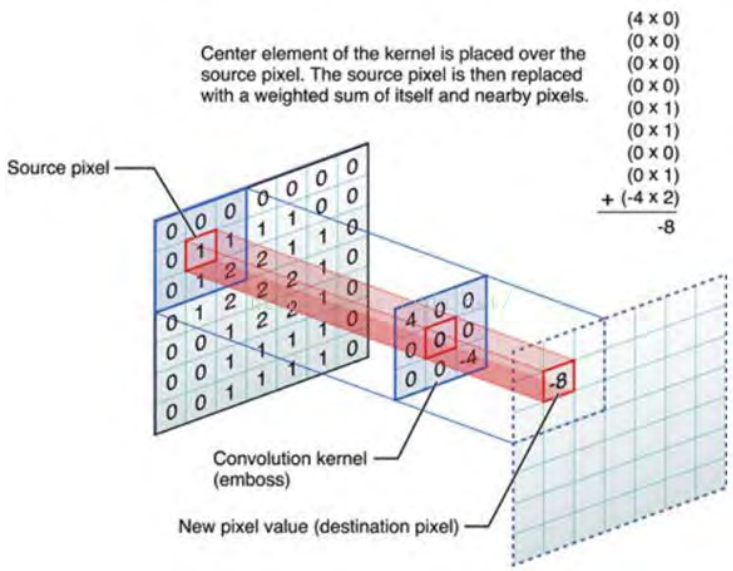


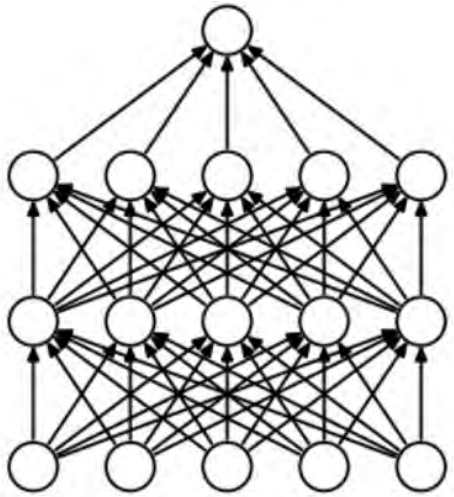
What we see

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

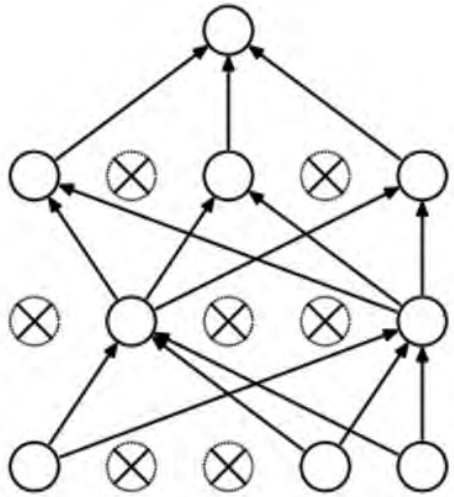
What a computer sees

Basis

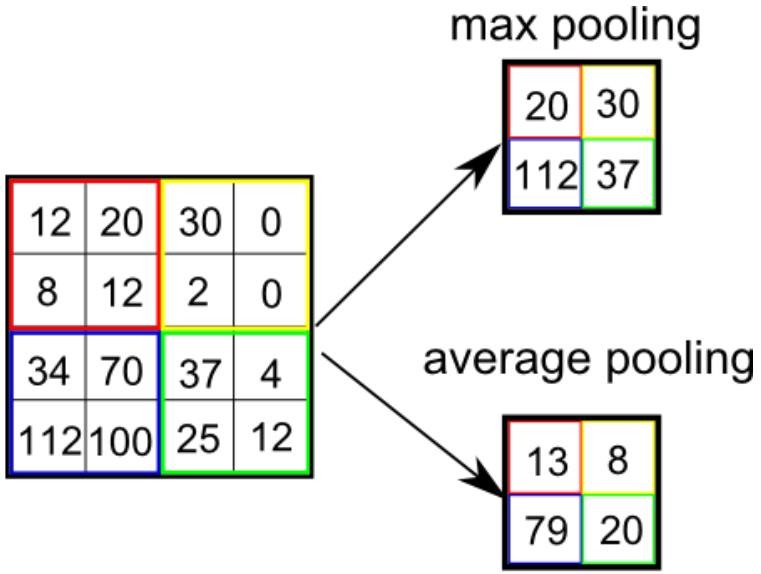




(a) Standard Neural Net



(b) After applying dropout.



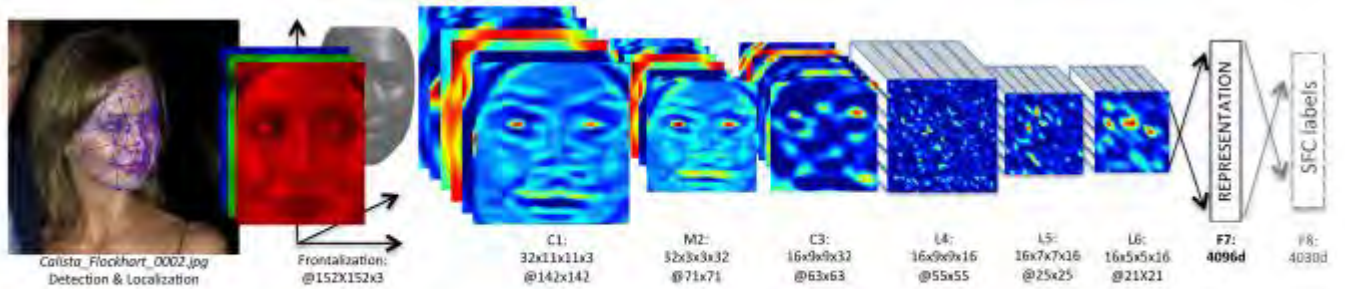
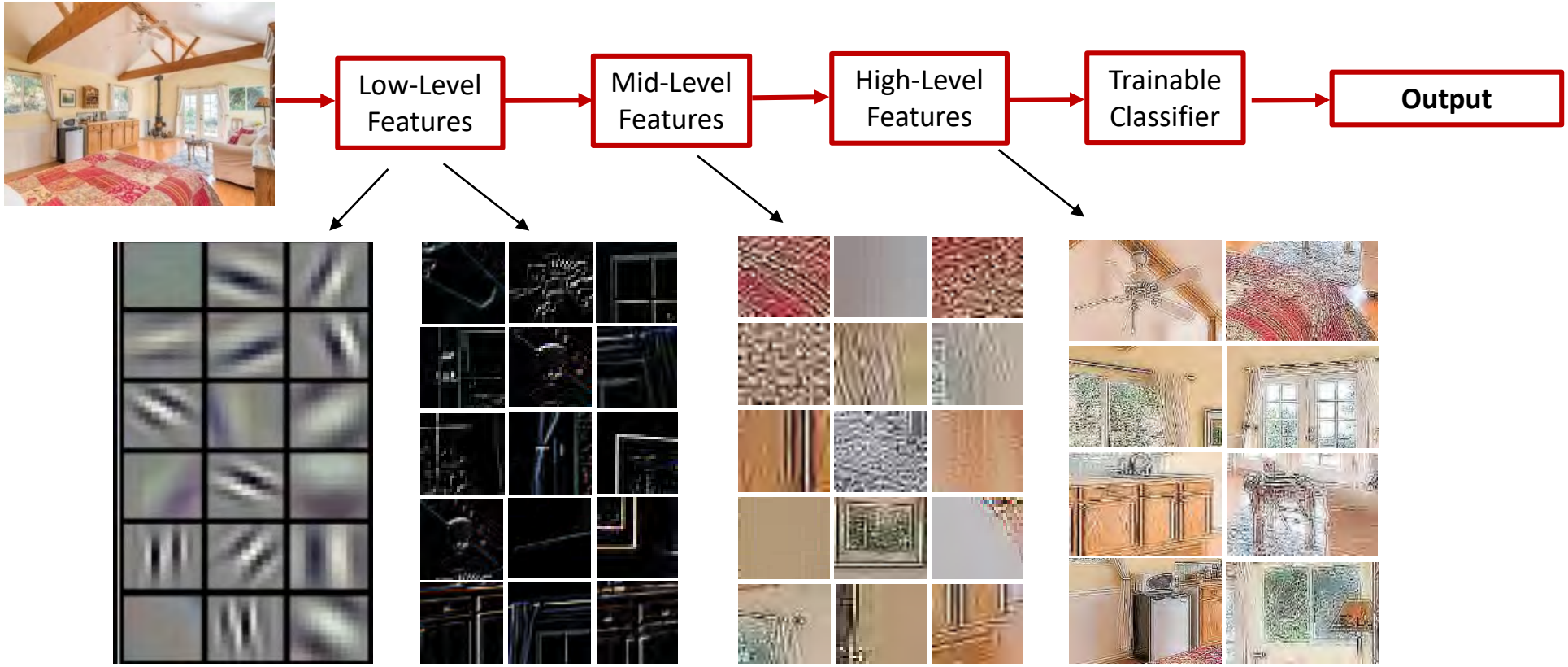


Figure 2. **Outline of the *DeepFace* architecture.** A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

ML vs. Deep Learning

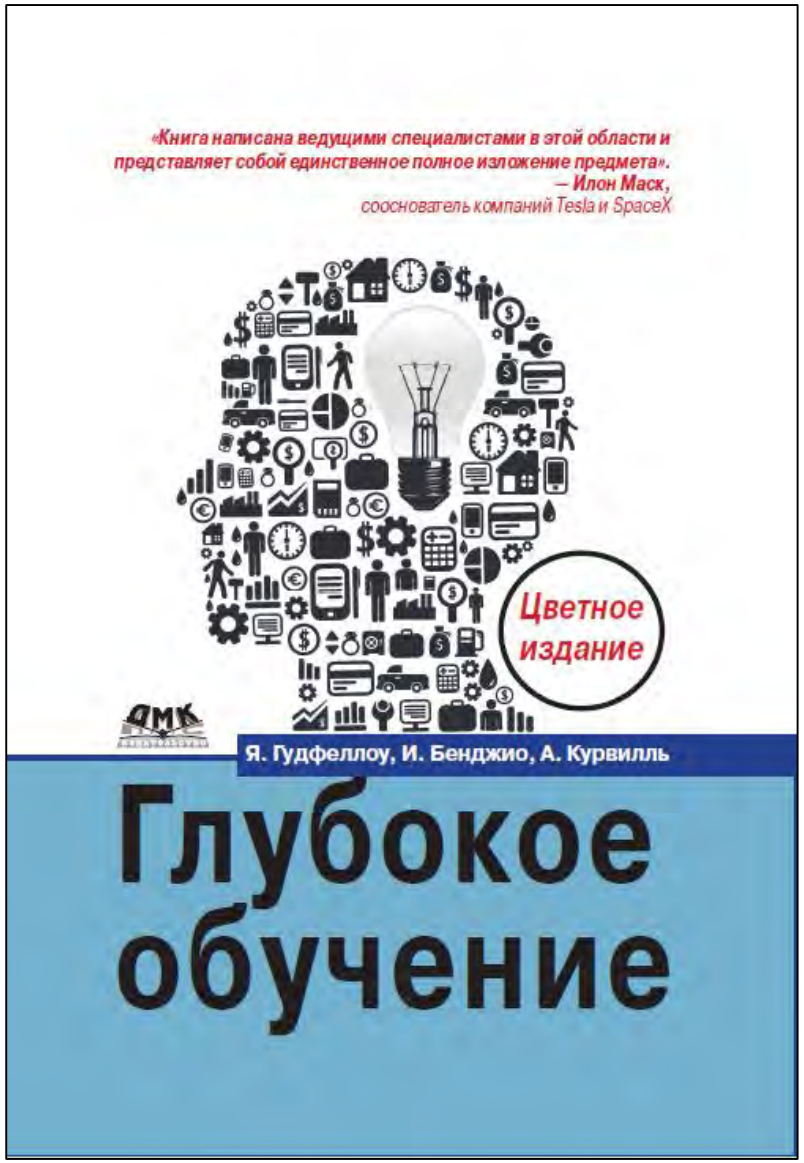
DL applies a multi-layer process for learning rich hierarchical features (i.e., data representations)

Input image pixels → Edges → Textures → Parts → Objects



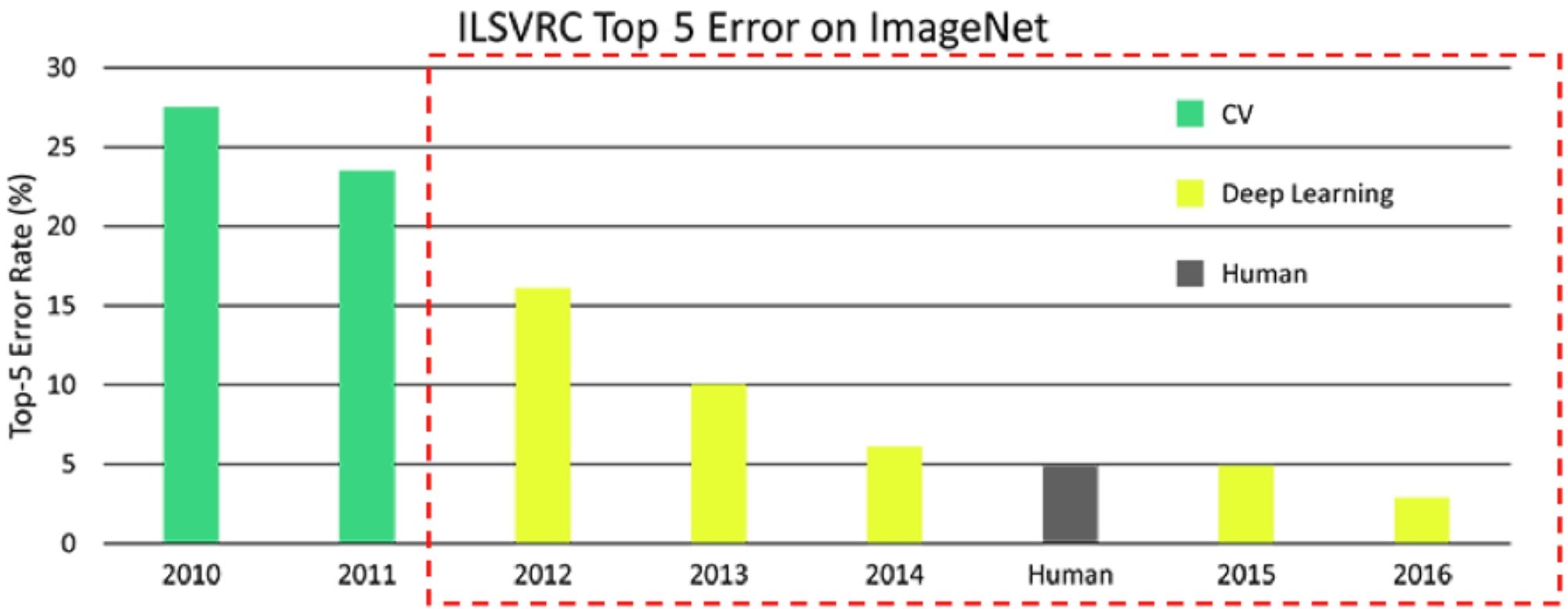
Basis


O'REILLY
Программируем с PyTorch
Создание приложений глубокого обучения
Ян Пойнтер


«Книга написана ведущими специалистами в этой области и представляет собой единственное полное изложение предмета».
— Илон Маск,
сооснователь компаний Tesla и SpaceX
Цветное издание
ДМК
Я. Гудфеллоу, И. Бенджио, А. Курвилль
Глубокое обучение

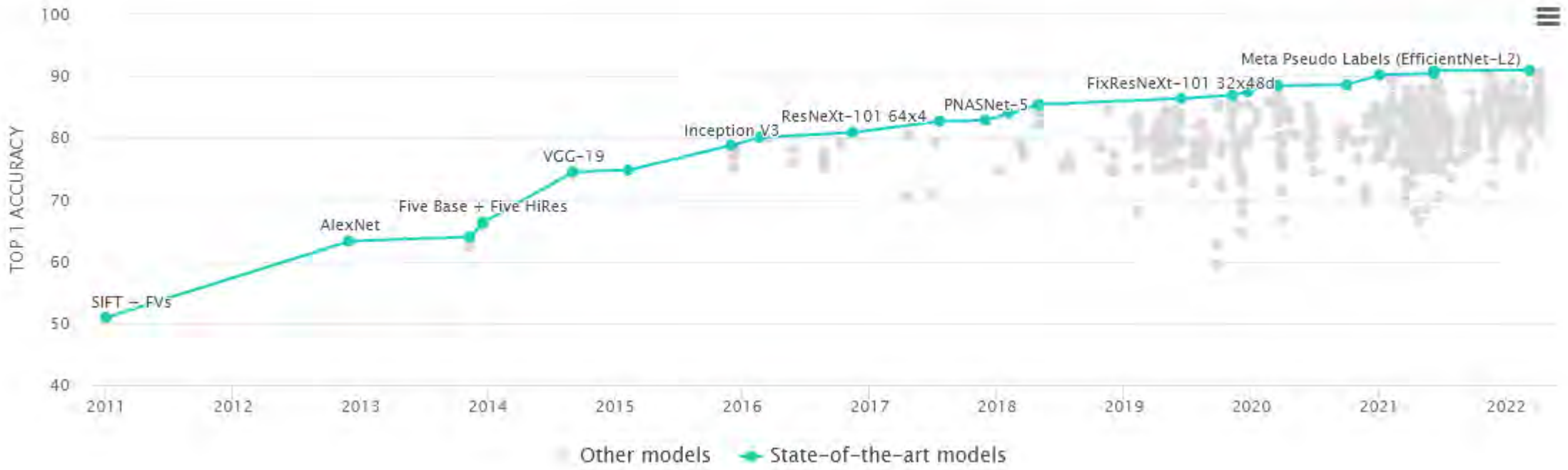
Basis

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)



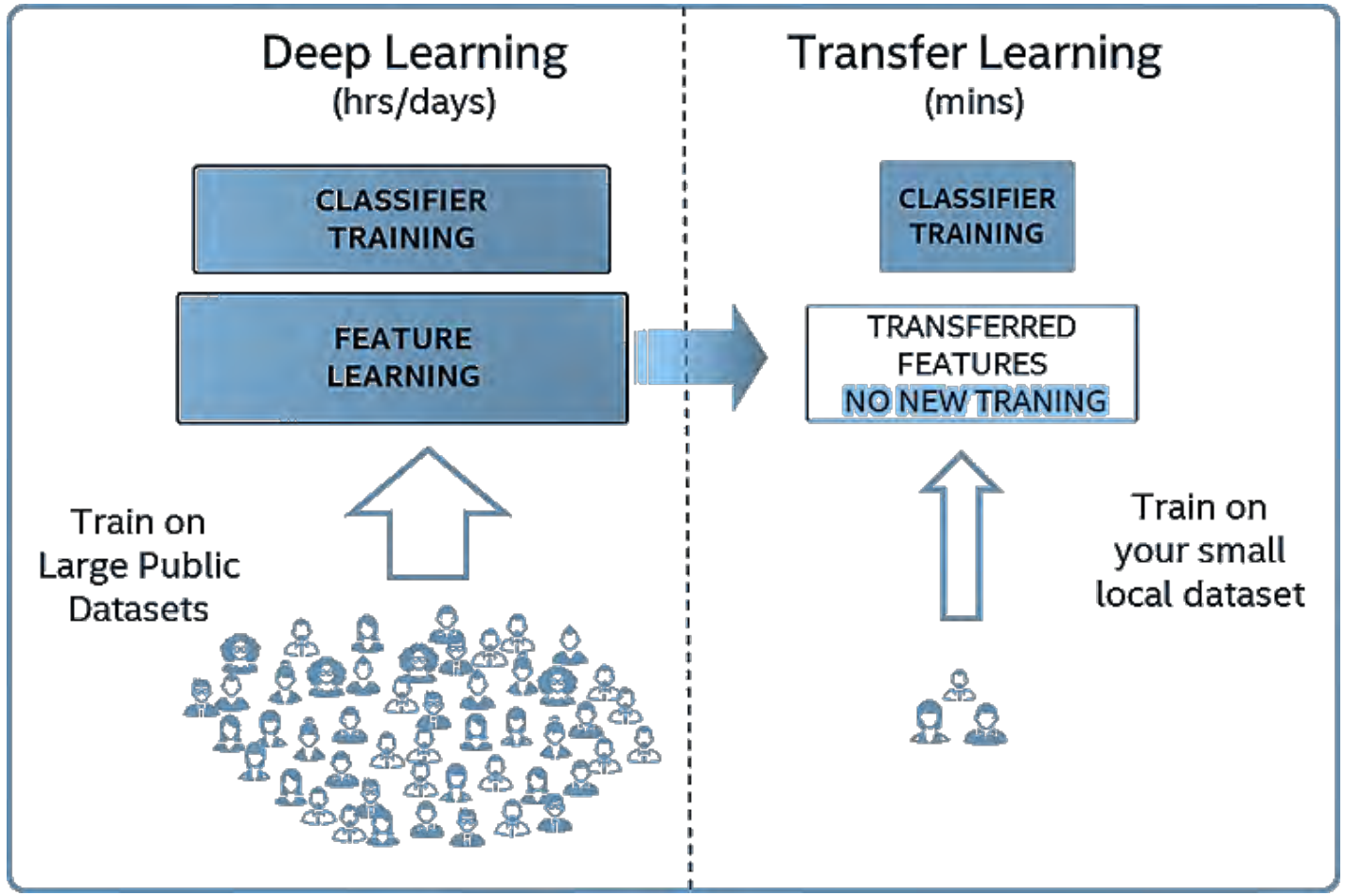
The introduction of Deep Learning techniques drove performance on image categorization from 30% error rates in 2010, down to <2% in 2017

Basis


100
90
80
70
60
50
40
TOP 1 ACCURACY
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
SIFT + FVs
AlexNet
Five Base + Five HiRes
VGG-19
Inception V3
ResNeXt-101 64x4
PNASNet-5
FixResNeXt-101 32x48d
Meta Pseudo Labels (EfficientNet-L2)
Other models
State-of-the-art models

Basis

How to learn if lack of data
Transfer Learning



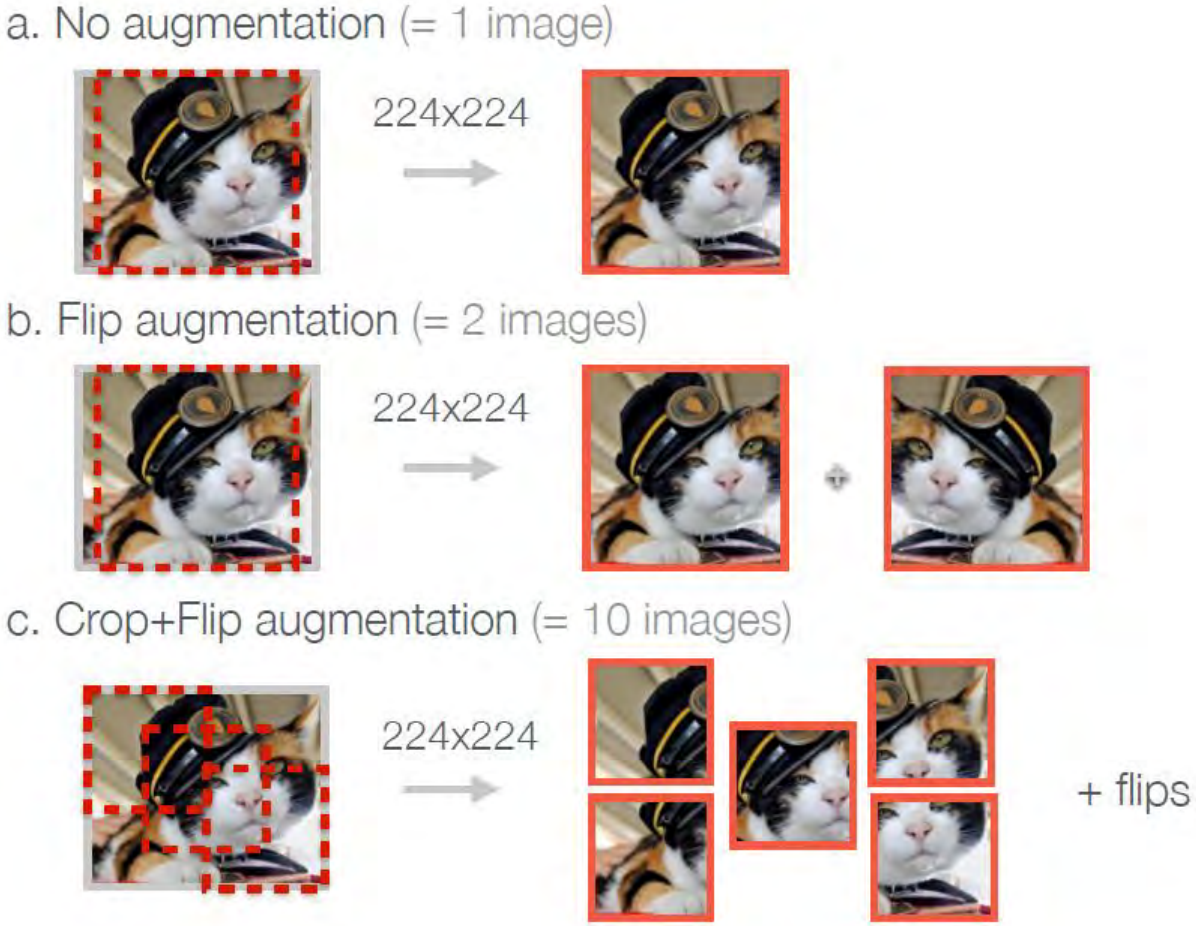
- Find a deep neural network pretrained on a big dataset
- Replace the classification layer with a layer appropriate for your task
- Finetune the new classifier on specific data
- Voila! Use the new model for inference

Autoaugmentation

Random Search learner This learner is a purely randomised searcher.

The genetic learner has similar elements to the Random Search learner, but uses information from previous sub-policies when generating new ones to more efficiently search for optimal augmentation parameters.

GRU with PPO updating Agent (gru learner.py) An GRU controller was used, which output a policy in the form of a length 10 sequence of vectors, each vector representing a operation. The GRU controller was updated using proxima policy optimization(PPO), using the accuracy of the child network as the reward value. In the context of the PPO update, which was developed in the reinforcement learning literature, the subpolicies are the 'actions' of RL agent.



One-shot / Few-shot Learning:

Few-shot learning (FSL) is a branch of supervised machine learning, focused on learning from limited number of examples[1]

Inspired by human/animal ability to rapidly generalize from few examples.

Typical scenarios[1]:

- Learning for rare cases – when obtaining labelled data is hard or impossible
- Reducing data gathering effort and computational cost

Variations:

- (<50, e.g. 5,10)-shot learning: General case, where only few examples are given
- 1-shot learning: Extreme case, where only one data example is given (e.g. 1 image per class)
- 0-shot learning: Instead of image, we have description of new class
- 'Less than 1'-shot learning[2]: N classes, M examples, M<N, use soft-labels

Why does it work?

In ML, experience is gained by fitting data.

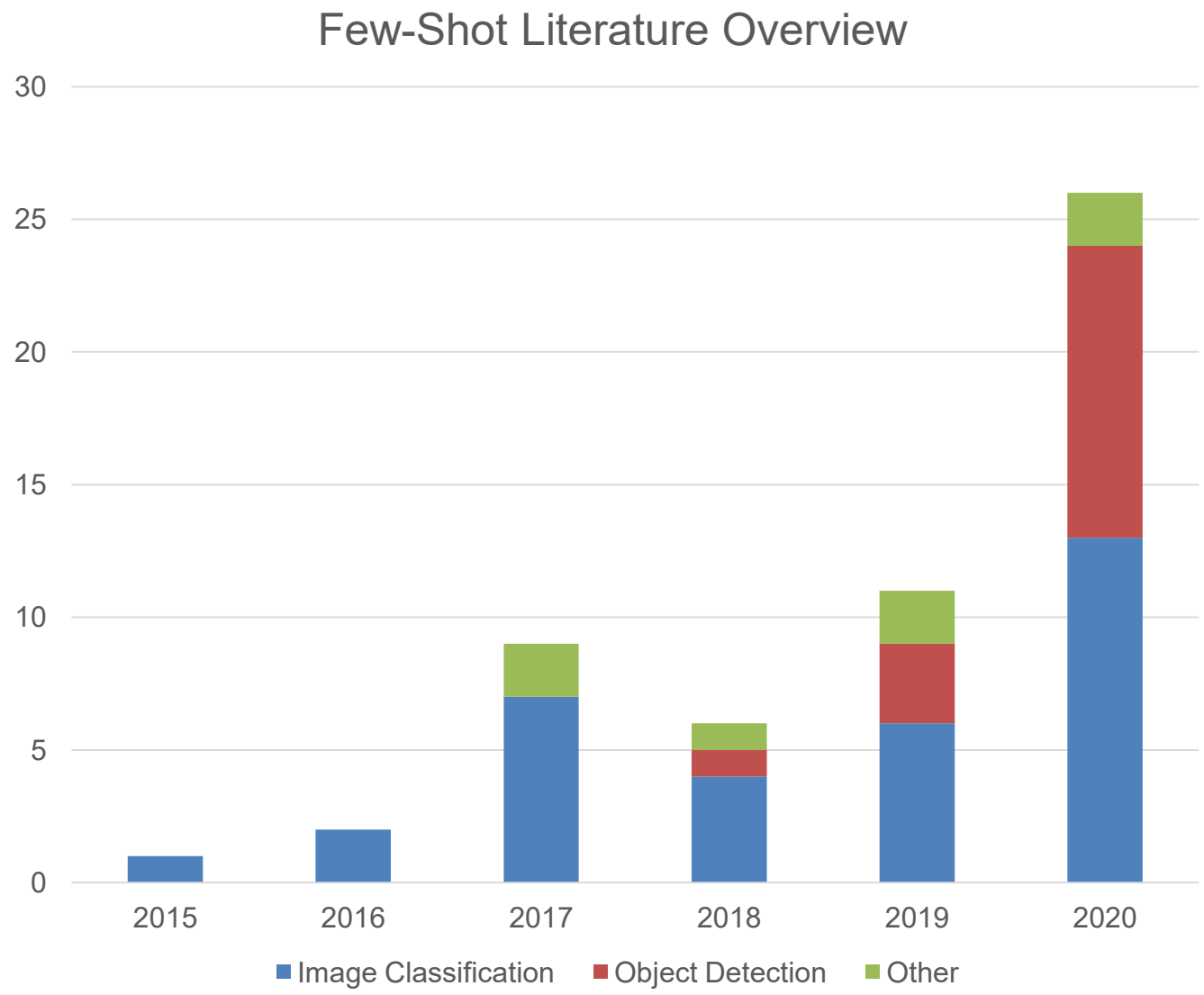
In FSL we also have prior knowledge[1]. Two training phases:

1. Gain prior knowledge by fitting large amount of examples that are similar to goal task
2. Gain experience on small dataset for goal task.

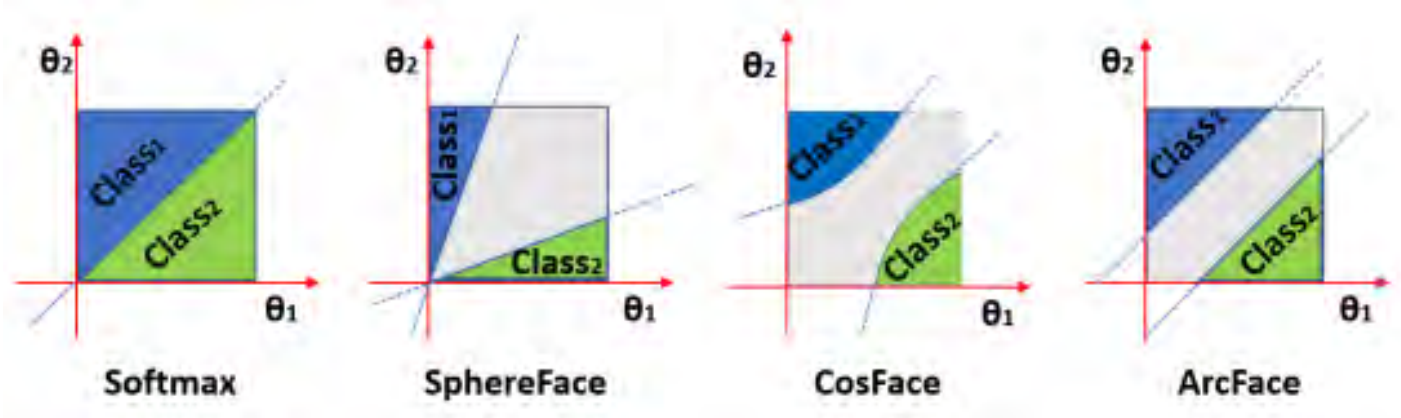
Few-shot learning literature:

- A lot of image classification papers
  (paperswithcode.com show 69 papers)
- In 2018 FSL starts object detection
  (paperswithcode.com show 12 papers, we can find more)
- Other topics include image generation, NLP, regression, etc.

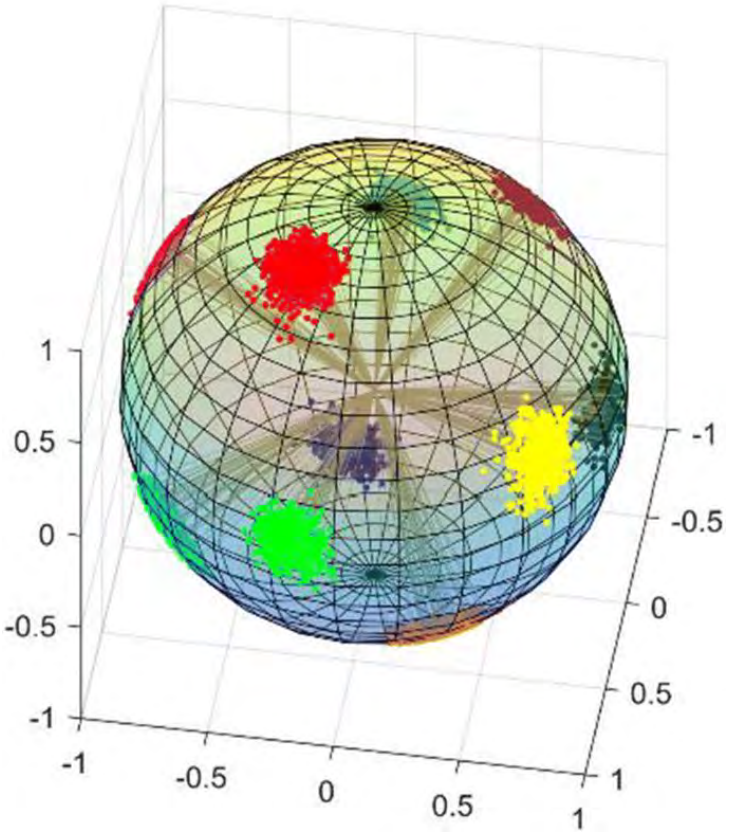
Note: 2018 is undersampled



Geometrical loss functions: arcface sphereface cosface

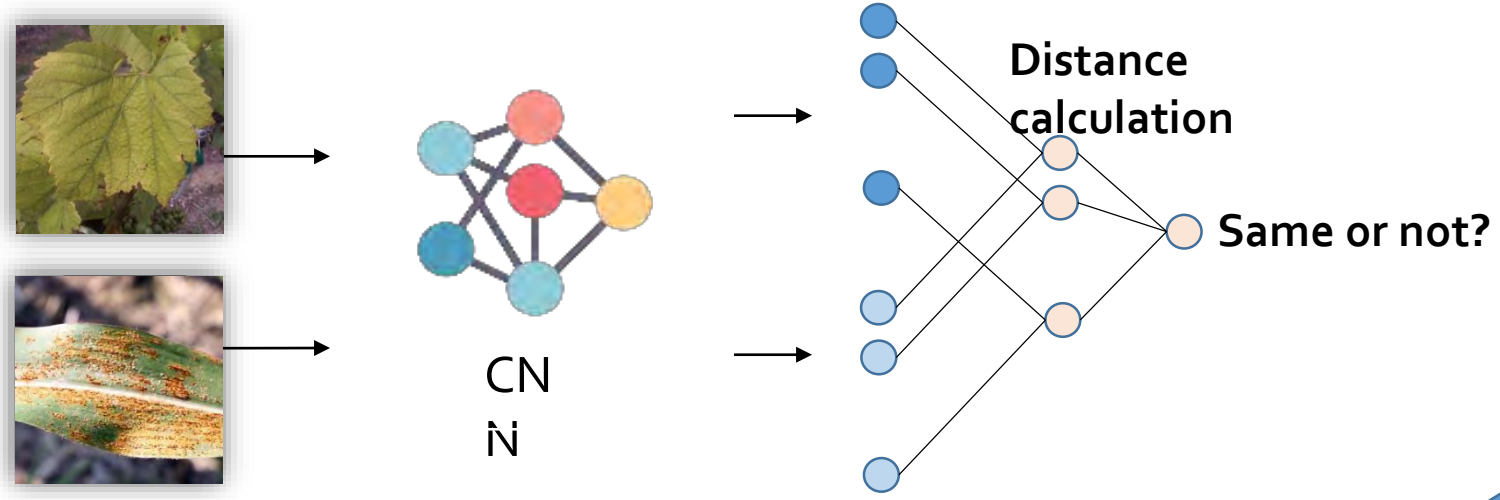


ArcFace, or **Additive Angular Margin Loss**, is a loss function used in face recognition tasks. The softmax is traditionally used in these tasks. However, the softmax loss function does not explicitly optimise the feature embedding to enforce higher similarity for intraclass samples and diversity for inter-class samples, which results in a performance gap for deep face recognition under large intra-class appearance variations.



How to learn if lack of data - Siamese Networks

Siamese networks is a part of **one-shot learning** approach. One shot-learning aims to learn information about object categories from one, or only a few, training samples/images



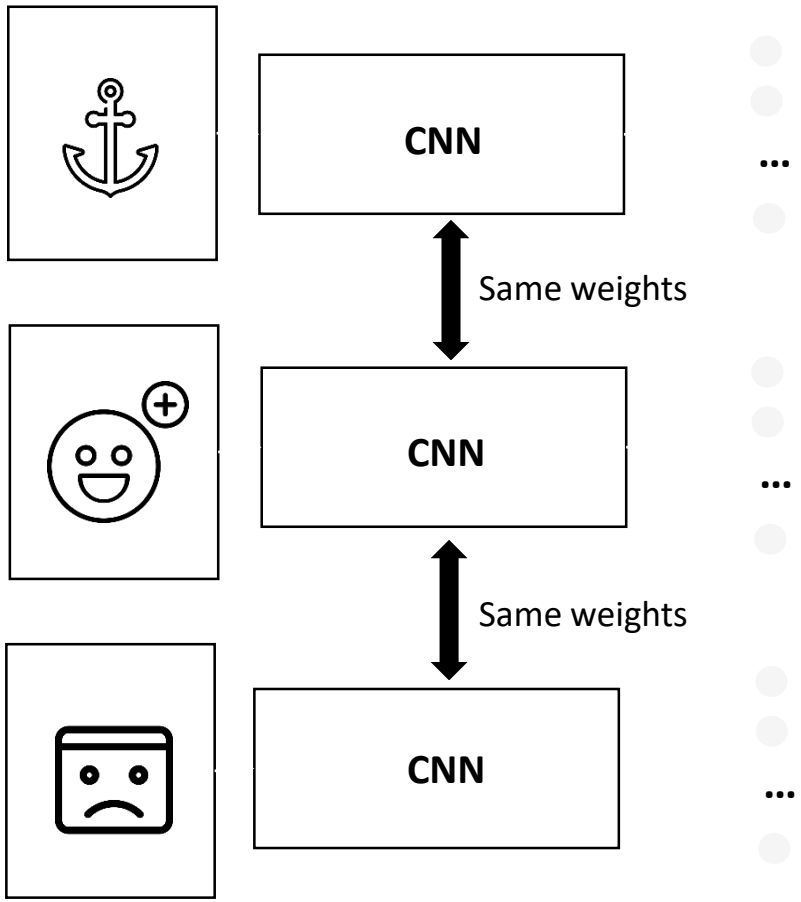
- Train siamese network on your own paired data

- Take the trained twin and append a classifier on top of it
- Finetune the new classifier on specific data appropriate for your task

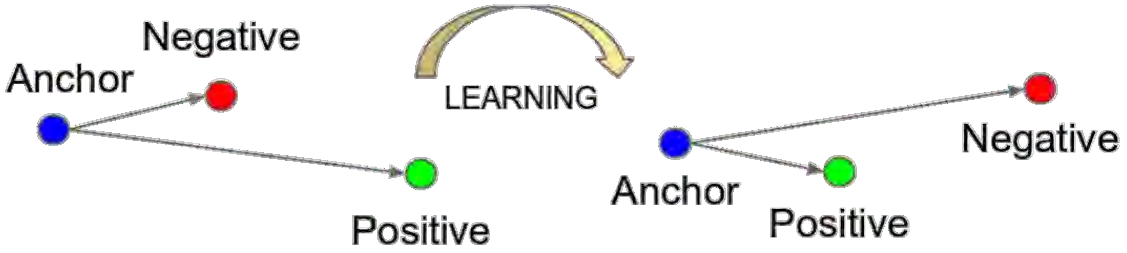
Same as for transfer learning

Basis

How to learn if lack of data
Triplet Networks



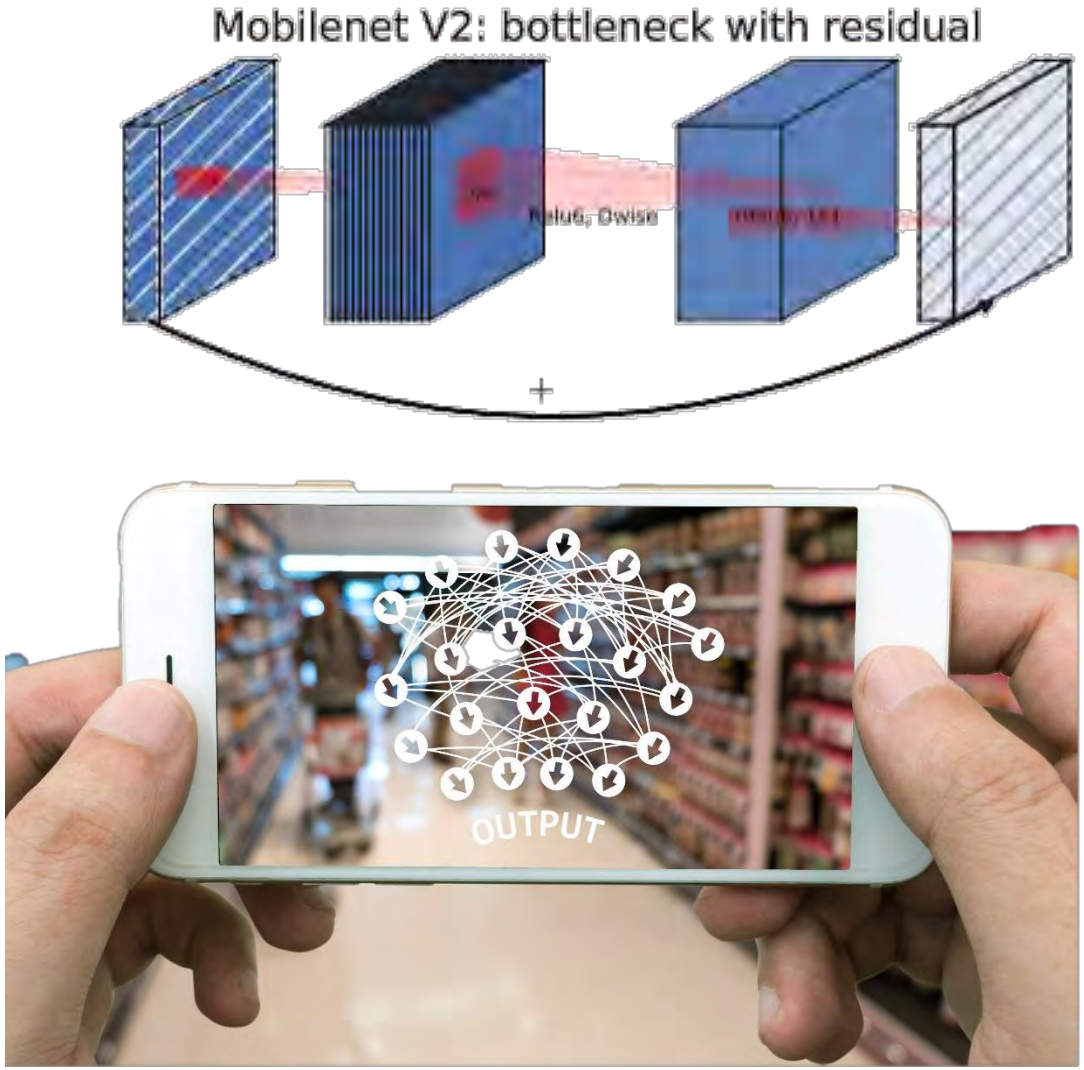
Triplet loss

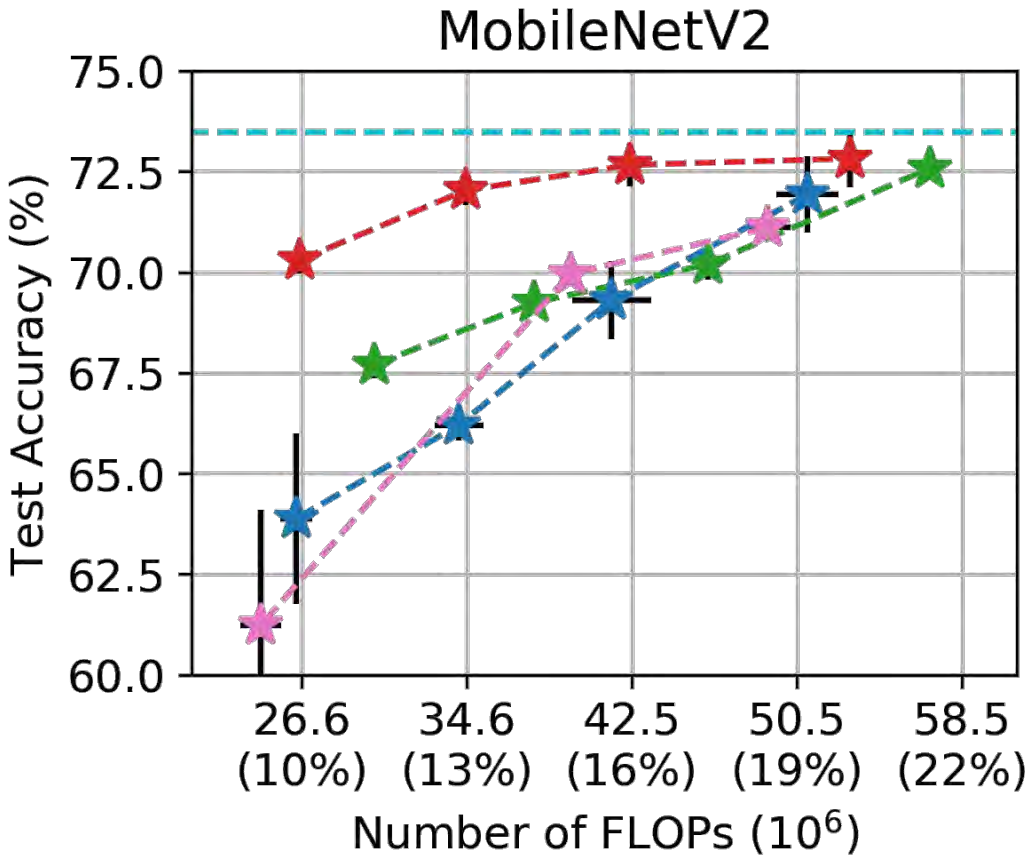


$$L = \max(d(a,p) - d(a,n) + margin, 0)$$

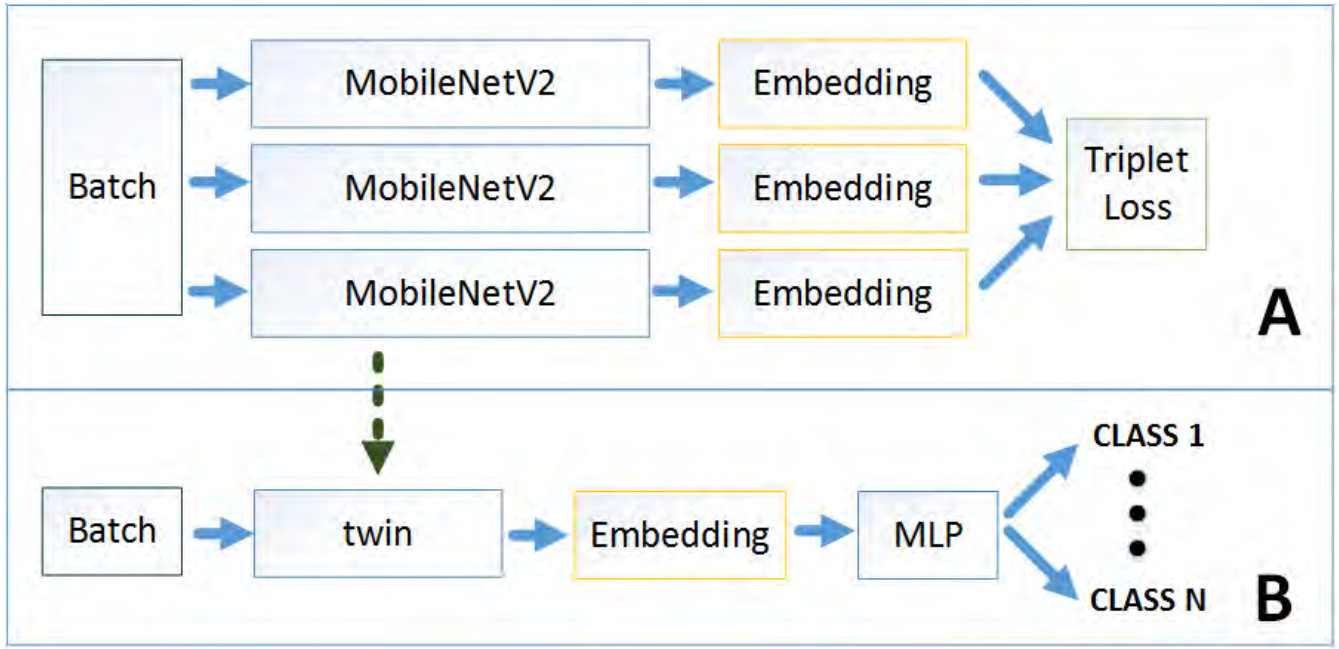
- **"d"** is some kind of function for calculating the distance between vectors, for example, Euclidean distance.
- A — **"a"** is an anchor image which we want to identify
- P — **"p"** image the same class as anchor
- N — **"n"** image of another class not matching the anchor

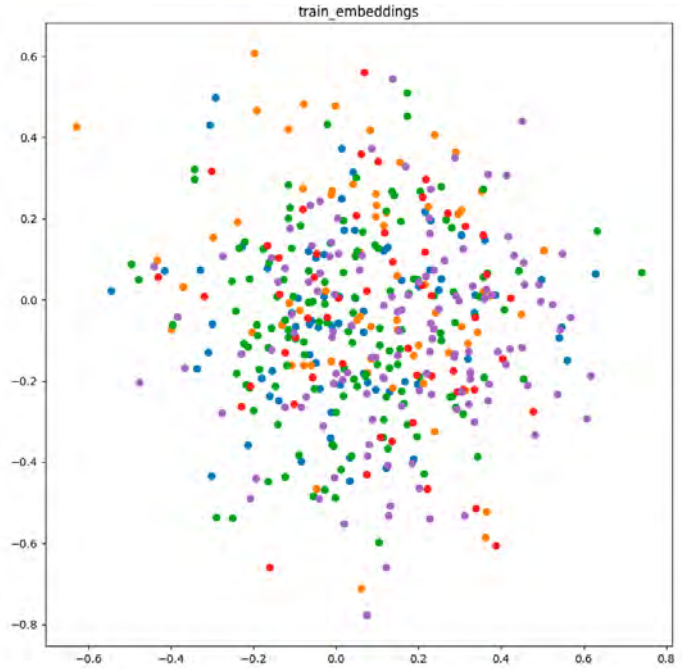
Deep Learning optimized for smartphones

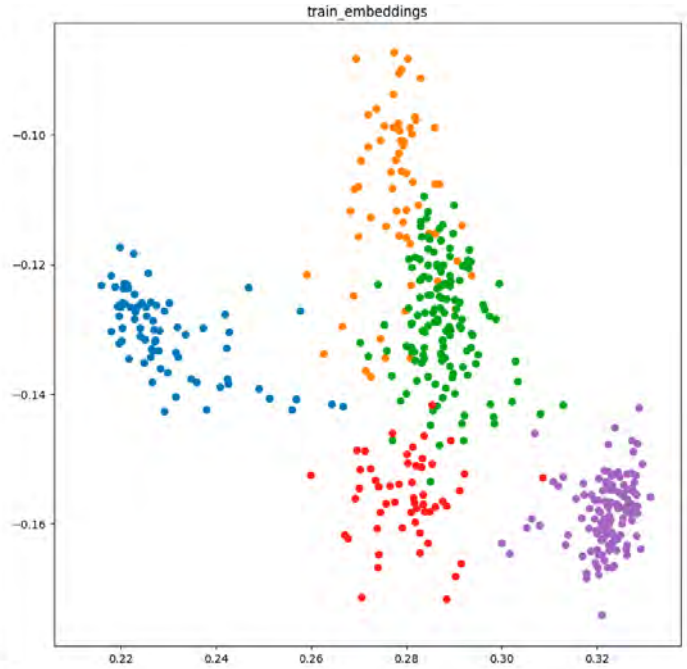




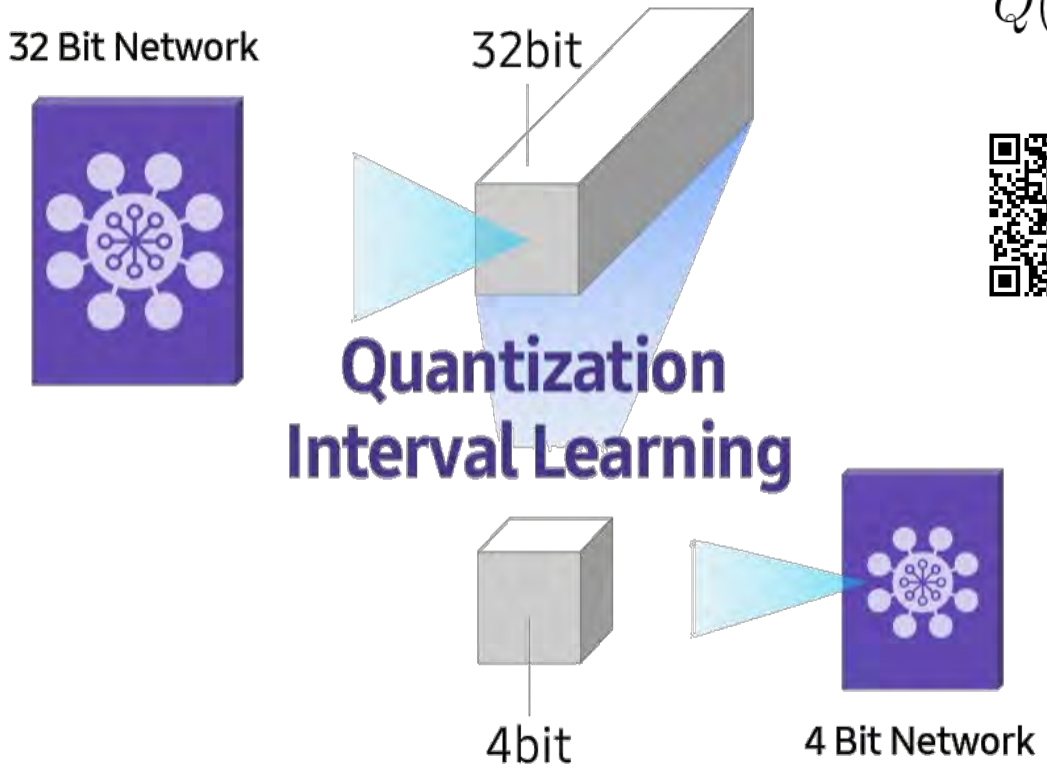
General pipeline







Quantization: smaller, faster, better?

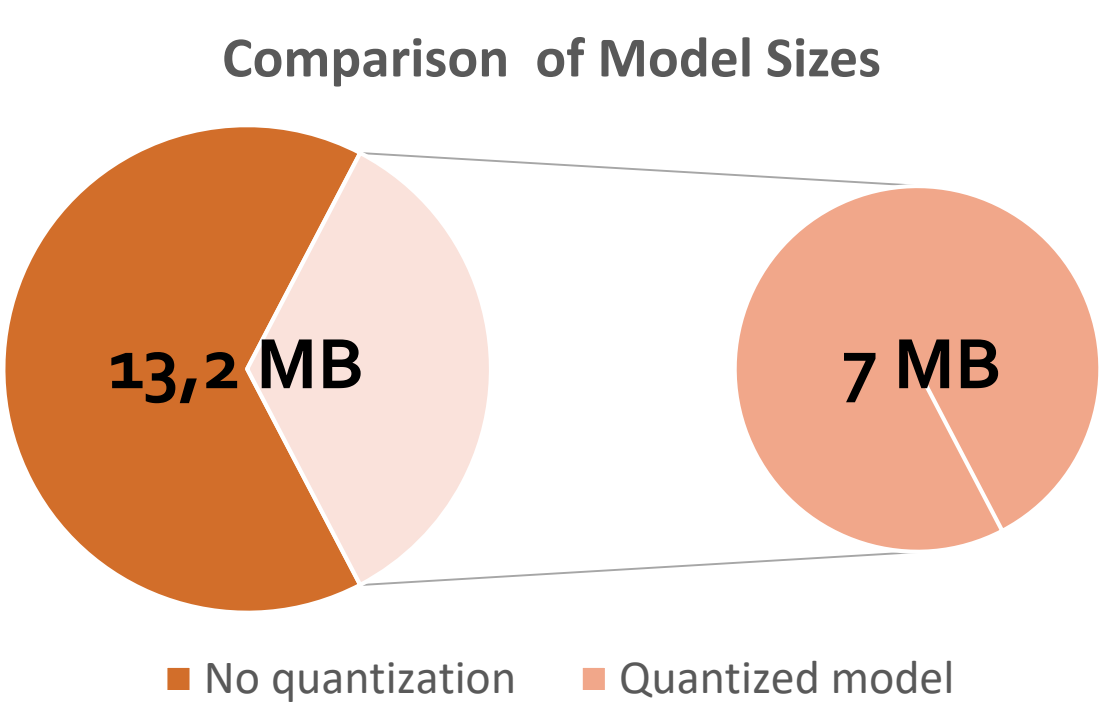


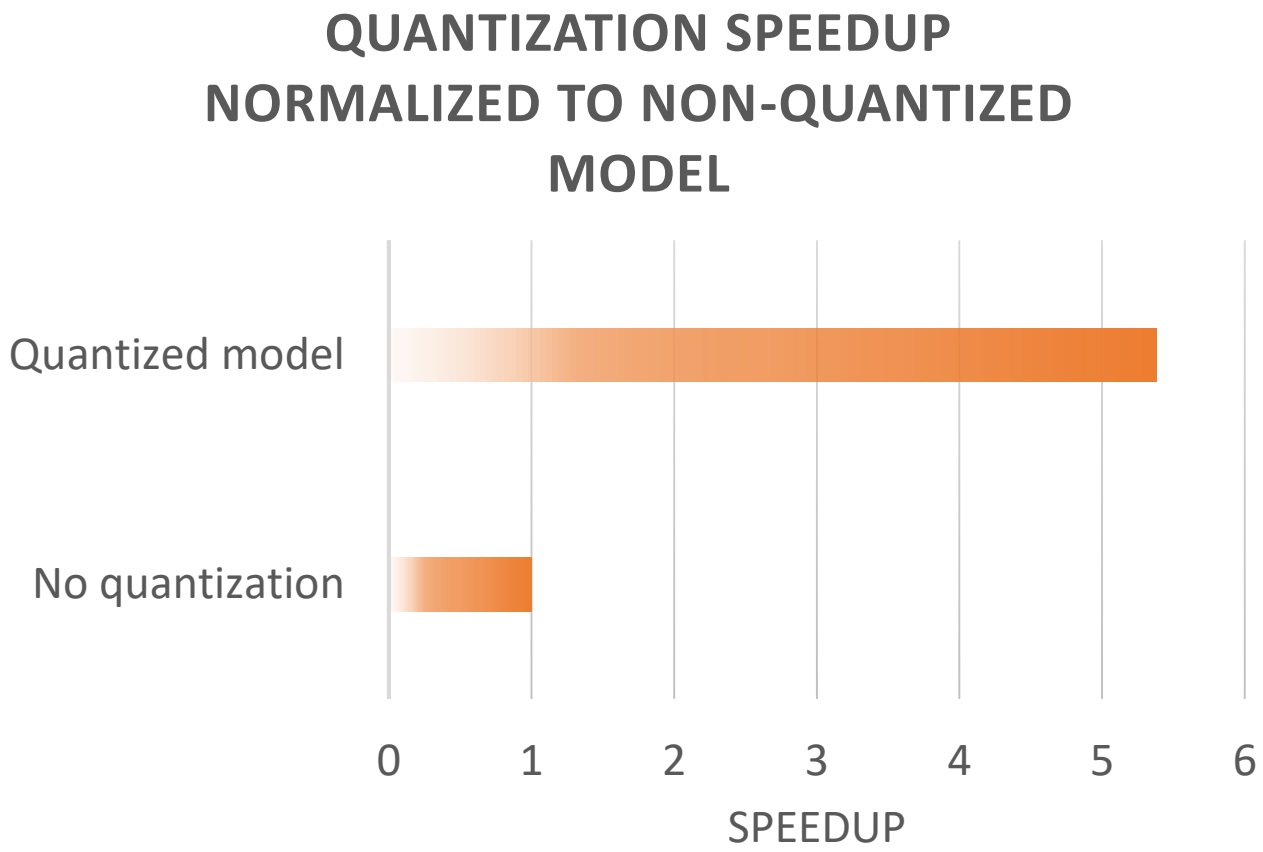
$$Q(x, \text{scale}, \text{zero\_point}) = \text{round}\left(\frac{x}{\text{scale}} + \text{zero\_point}\right)$$

Quantization helps speed up inference on devices:

- x86 CPUs with AVX2 support or higher (without AVX2 some operations have inefficient implementations)
- ARM CPUs (typically found in mobile/embedded devices)

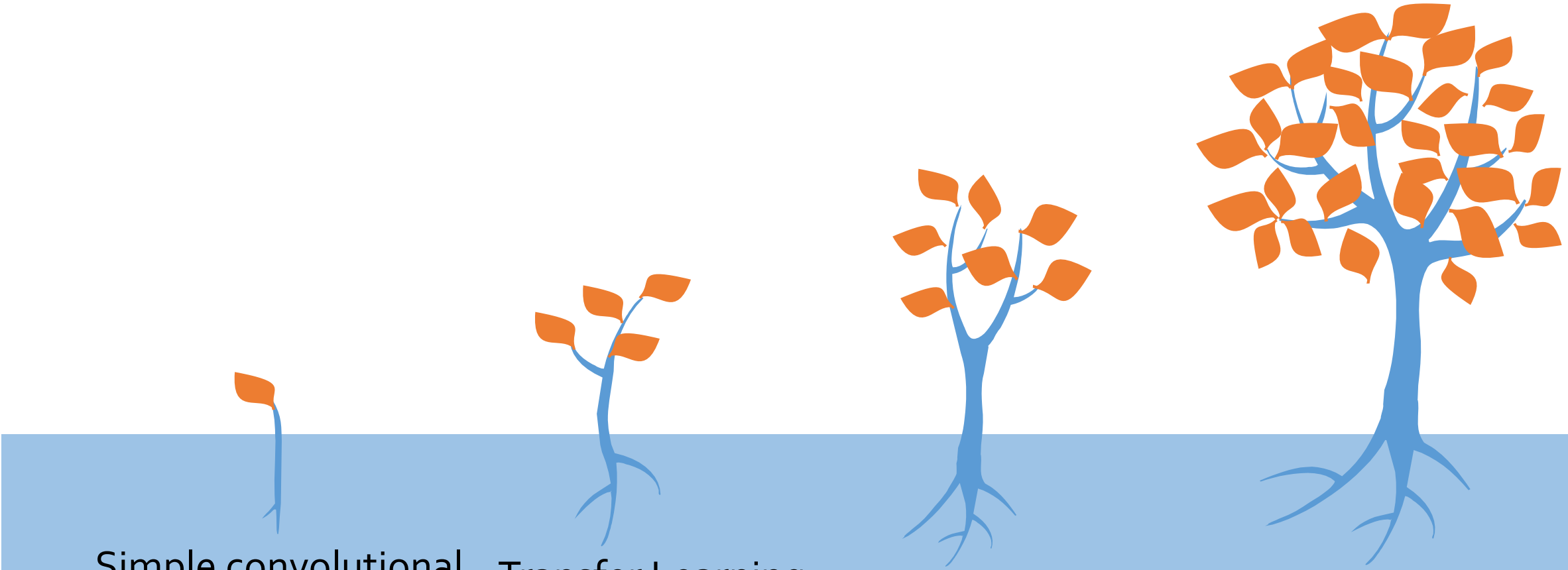
Quantization Results





Accuracy remains the same!

Evaluation of the Results



Simple convolutional neural network

Accuracy less than 50%

Transfer Learning model

Accuracy less than 80%

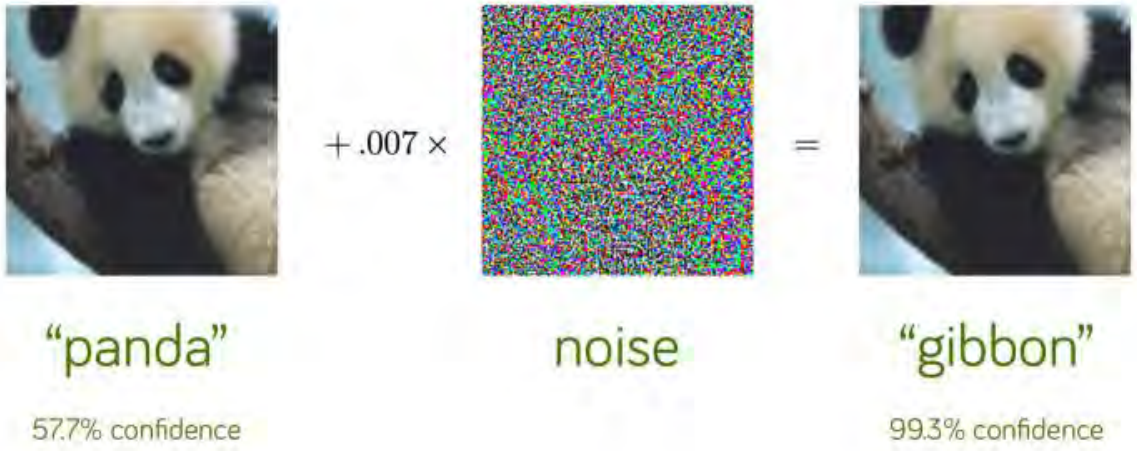
Siamese network

Accuracy less than 90%

Triplet network

Accuracy ~98%

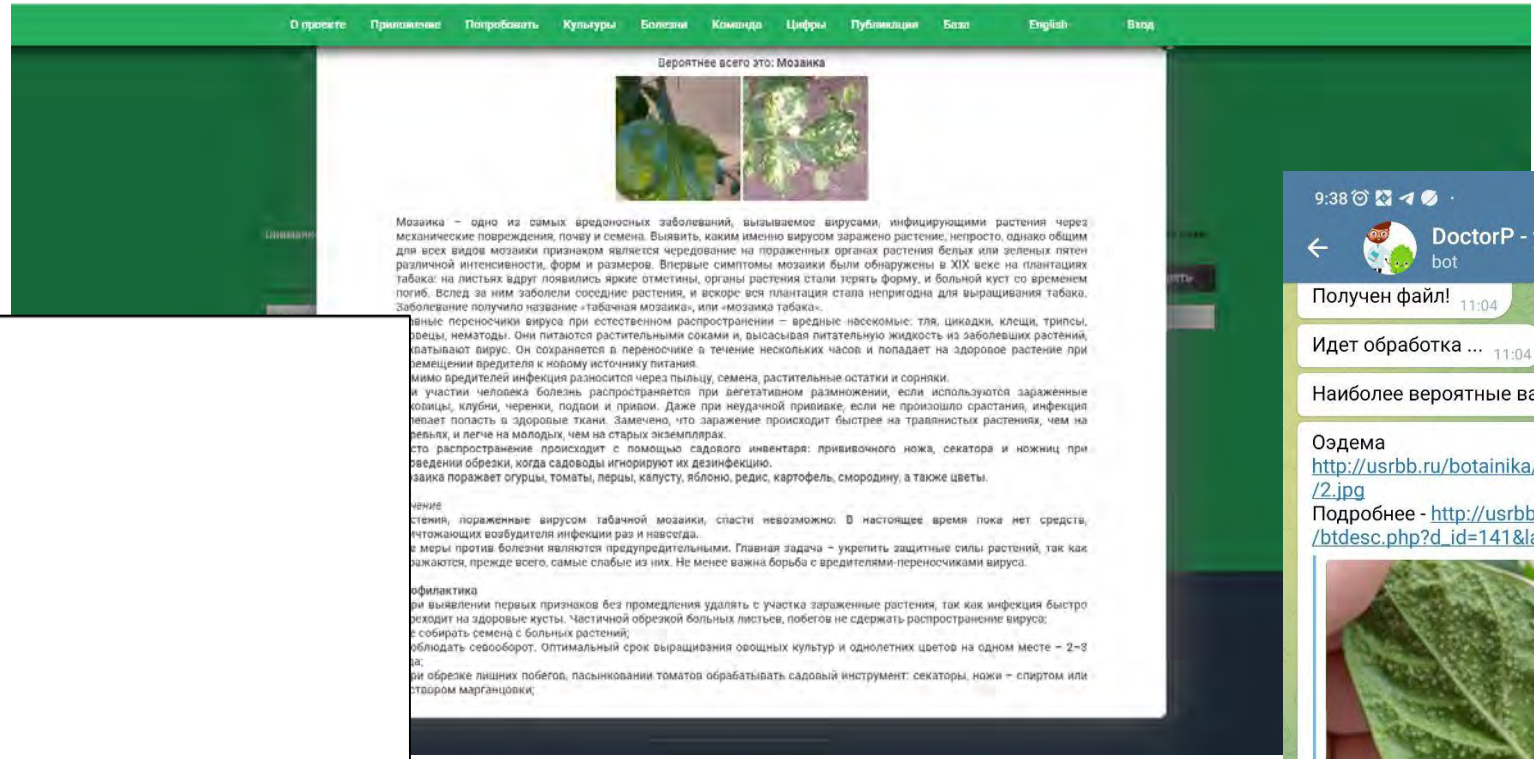
Attacks on Deep Learning Visual Classification



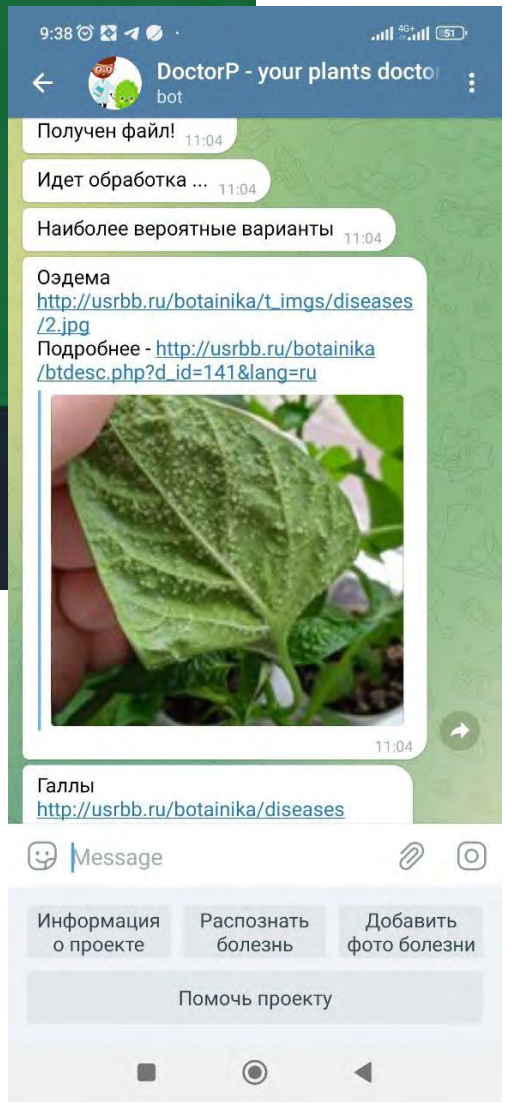
The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows a physical perturbation applied to a Stop sign. The systems classify the sign on the right as a Speed Limit: 45 mph sign!

Реализция

PDDP web-portal, telegram-bot, API, app

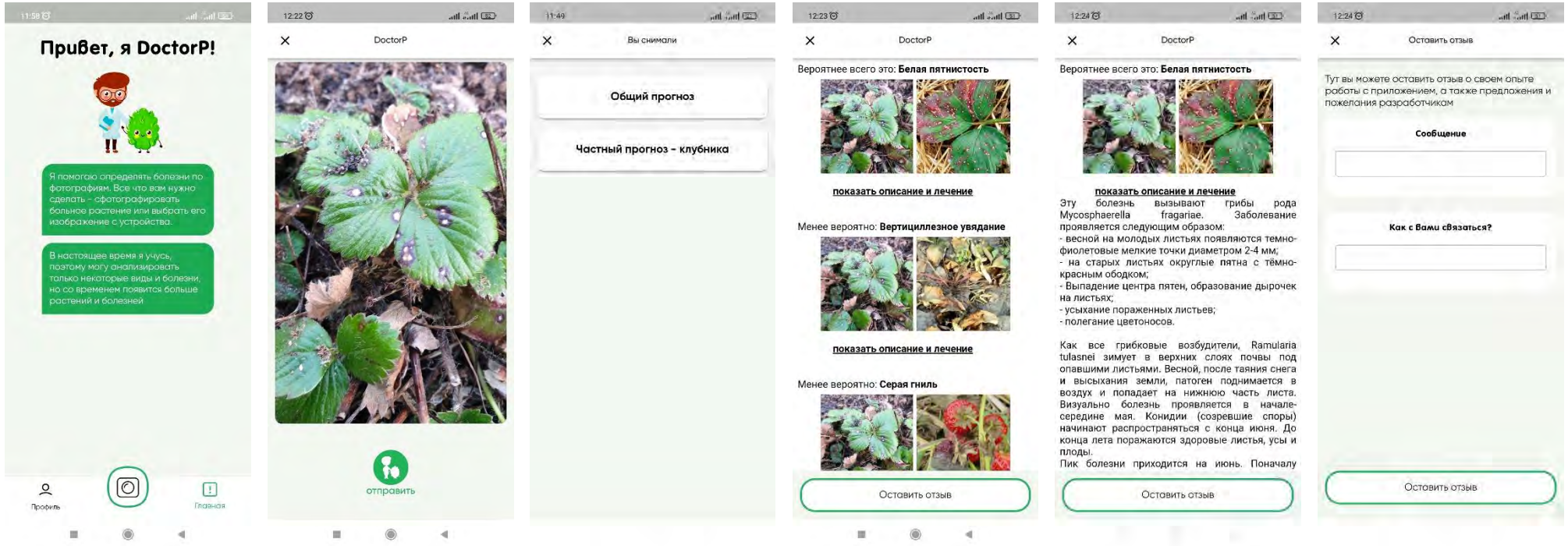



```
{
  "error":0,
  "general_predictions":[
    {...},
    {...},
    {...}
  ],
  "custom_predictions":[
    {...},
    {
      "type":"розы",
      "prediction":[
        {
          "name":"Мозаика",
          "sample":"http://usrbb.ru/botainika/diseases/rmv1.jpg",
          "description":"<p>Это наиболее распространенное заболевание вирусного типа. Узнать его очень просто – листья покрываются хлоротичными пятнами и узорами, которые могут перейти в прожилковый хлороз. Листья деформируются (узколистность, курчавость, морщинистость) и постепенно опадают. Рост побегов замедляется, и они не вызревают. Побеги, которые больны, вырезают. При дальнейшем распространении болезни растение уничтожают
```

PDDP mobile App

Google play: “**DoctorP**”. * Only android version is available



The user has the opportunity to photo a diseased plant and get a prediction for the disease and treatment suggestions. It is possible to download images in the absence of the ability to take a photo. The application requires access to the Internet to work.

We can run the model on the mobile device directly (we have tried it) but models changing too often.

Server side

Здравствуйте, Александр Ужинский

Домой | Выход

Пользователи | Культуры/Болезни | Описание/Лечение | Изображения | Запросы пользователей | Запросы от экспертов

Описание/Лечение:

Орхидеи - Серая гниль

Описание:

Paragraph

Обычно первые признаки серой гнили появляются на лепестках растения и
водянистых пятнышек. Практически одновременно с ними, если внимательн
возникают различных размеров участки, сначала мелкие, затем они станов
сероватого оттенка. Они будто бы присыпаны золой, и если провести по ни
всяких сомнений, серая гниль на орхидее
Появляется так же у основания из-за повышенного уровня влажности, может быть ещё довольно сырой грунт, а также прохладный воздух в помещении, где произрастает орхидея. О том, что заболевание уже поразило растение, говорят пятна тёмного оттенка. Эти пятна покрываются серым налётом. Растение довольно быстро вянет.

Профилактика
• Необходим карантин для вновь приобретенных растений.
• Регулярный осмотр.
• Грибок серой гнили поражает ослабленные орхидеи, содержащиеся в неправильных условиях при недостаточном уходе. Осмотрите место для выращивания на наличие благоприятных условий для серой гнили. Поскольку этот патоген процветает и зимует на мертвых или умирающих растительных материалах, удалите остатки растений, опавшие цветы и листья для того, чтобы уменьшить возможность распространения грибка.
• Все водные процедуры в холодный период года желательно осуществлять с утра или в первой половине дня, чтобы к

It also appears at the base due to the increased level of humidity, there may still be quite damp soil, as well as cool air in the room where the orchid grows. The fact that the disease has already struck the plant is indicated by spots of a dark shade. These spots are covered with a gray coating. The plant wilts pretty quickly.

Prevention
• Quarantine is required for newly acquired plants.
• Regular inspection.
• Gray mold fungus infects weakened orchids kept in the wrong conditions with insufficient care. Inspect the growing area for favorable conditions for gray mold. Because this pathogen thrives and overwinters on dead or dying plant material, remove plant debris, fallen flowers, and leaves to reduce the possibility of fungus spread.
• It is advisable to carry out all water procedures in the cold season in the morning or in the first half of the day, so that by the evening the plants have had enough time to dry out.

N:30455 (2022-11-15 09:30:00) PDD

Прогноз модели: Трипс|Тля|Паутинный клещ (фиалка:Фузариоз|Фитофтороз|Бактериоз) (розы:Паутинный клещ|Нехватка элементов|Серая гниль)
{"success": "true", "cropse1":"violet", "cropse2":"roses", "cropse3":"succulent", "full1":"Thrips", "full2":"Aphid", "full3":"Spider mite", "cropse1_disease1":"Fusarium", "cropse1_disease2":"Late blight", "cropse1_disease3":"Bacteriosis", "cropse2_disease1":"Spider mite", "cropse2_disease2":"Nutrient deficiency", "cropse2_disease3":"Botrytis blight"}

Определено верно
(Проверить)
Запросить добавление в базу
Требуется проверка

Культура (если модель определила не верно)
Болезнь (если модель определила не верно)

Запуски:0
Протоколы ухода:0
Первый клик:Частный прогноз - фиалка

Отправить

X

https://github.com/AndriyMulyar/semantic-text-similarity

TEXT CLASSIFICATION

A TRAINED BERT MODEL WAS USED TO CLASSIFY PLANT DISEASES. TEXT SUGGESTIONS ARE FED TO THE MODEL INPUT, AND THEY ARE CONVERTED TO VECTORS AT THE OUTPUT. THEN THESE VECTORS ARE COMPARED WITH VECTORS IN THE DATABASE OF TEXT DESCRIPTIONS OF DISEASES.

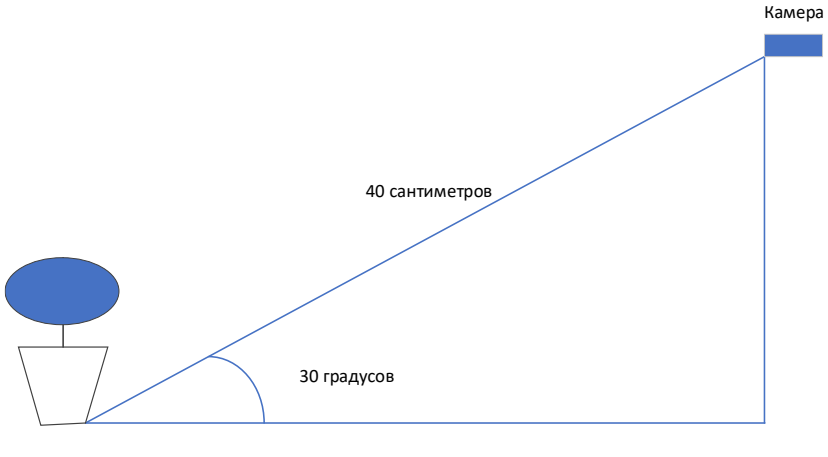
BERT MODEL

Задачи:

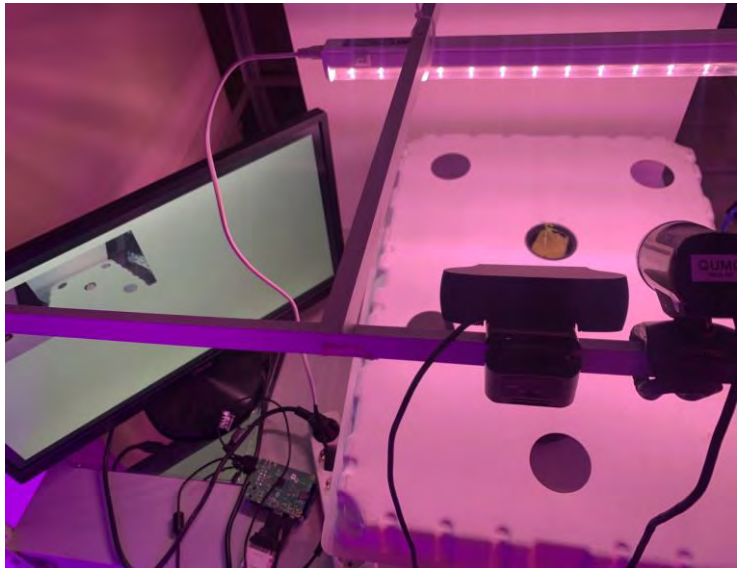
1 Расширение базы изображений
2. Эксперименты с различными функциями потерь (ArcFase, CosFace, etc)
3. Эксперименты с подбором оптимальных политик аугументации данных
4. Эксперименты с Unsupervised learning
5. Разработка автоматизированных инструментов поиска и пополнения базы изображений

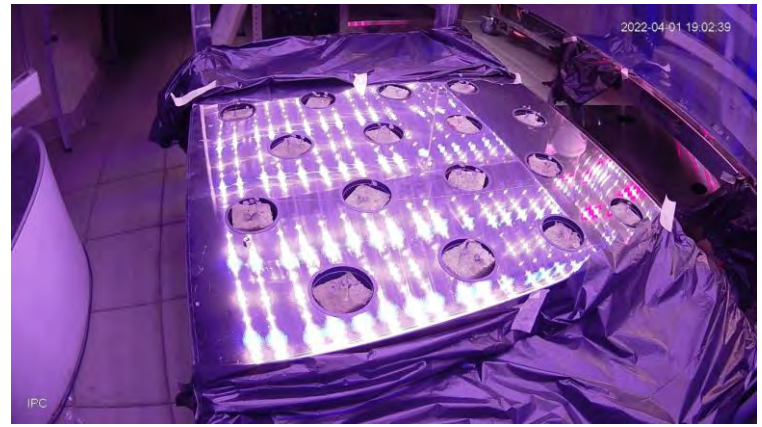
Отслеживание развития растений

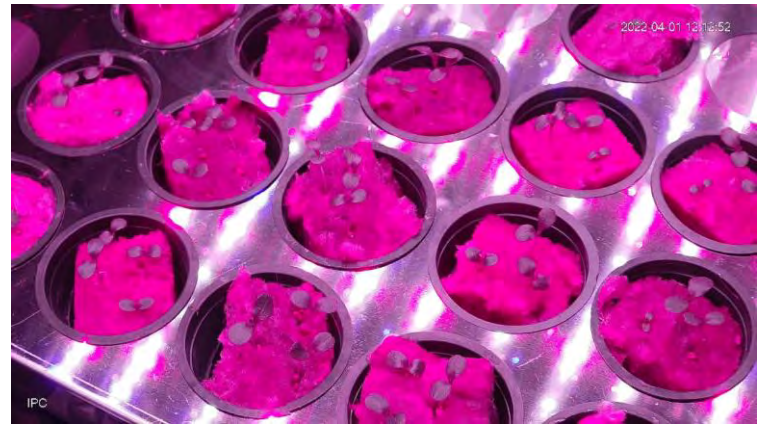
Совместный проект с Темирязевской академией в рамках проекта Научный центр мирового уровня “Агротехнологии будущего”



- **Классификация степени развития растения.**
- **Определение весовой группы растения.**







Другие проекты, но мы про них не расскажем - NDA

Салаты

Картошка

Яблоки

Сады

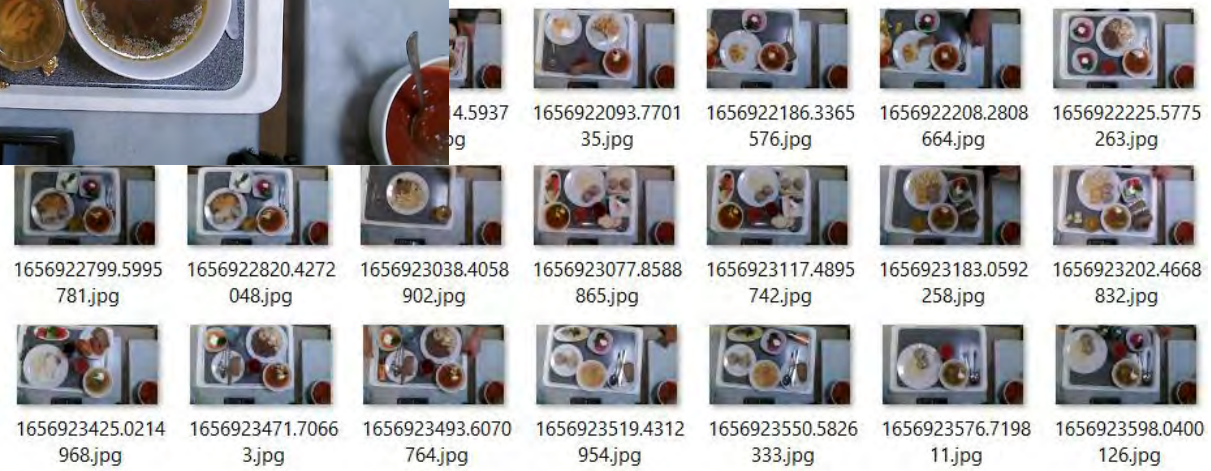
Set of images

Data collection was carried out in automatic mode using raspberry pi 4, a digital camera with manual focus, and sonar.

A set of images in 5 days:

- 446 images
- More than 150 different classes of objects

Dining room LIT JINR



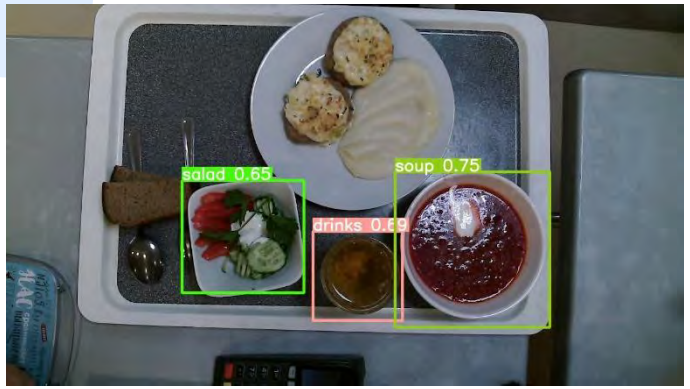
Experiment №1

Information:

Architectures: YOLOv5S
Batch size: 16
Epochs: 50
Number of classes: 7
Classes: 'bakery', 'drinks', 'garnire', 'main', 'other', 'salad','soup'

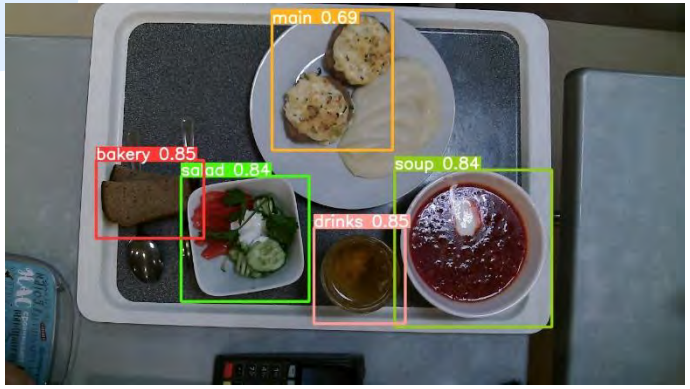
If there are not many classes (in our case 7), then yolo allows you to get good results, already with a training sample of 200 marked up images, and if there are more than 250 of them, then everything is finein general, but how much worse will it be if there are more image classes and the training sample is smaller?

Step #1



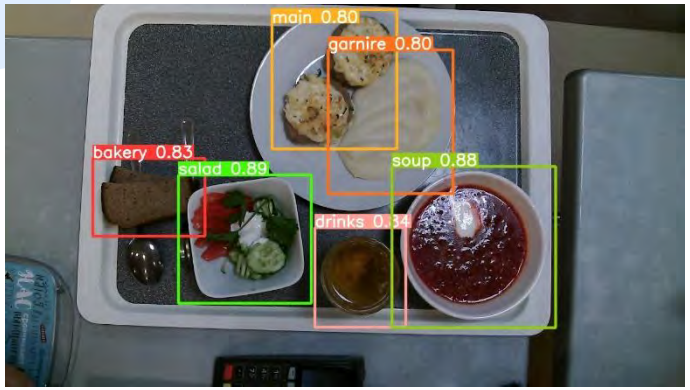
Recall: 0.666
Precision: 0.785
Map: 0.728
Count of images: 100

Step #2



Recall: 0.807
Precision: 0.924
Map: 0.904
Count of images: 200

Step #3



Recall: 0.926
Precision: 0.938
Map: 0.954
Count of images: 270

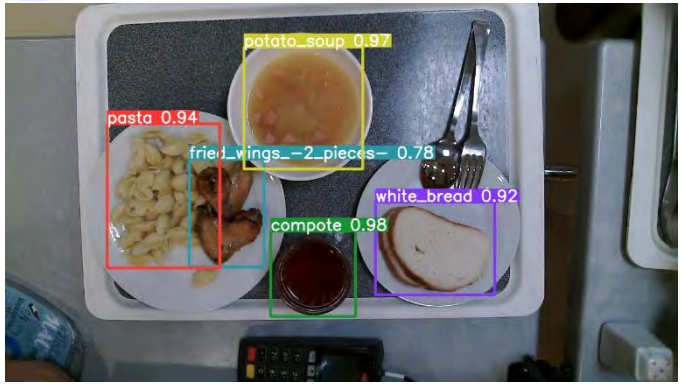
Experiment №2

Information:

Architectures: YOLOv5S, YOLOv6S, YOLOv7
Count of images: **126**
Batch size: 16
Epochs: **400**
Number of classes: 36 ('beetroot', 'black_and_white_salad_with_parsley', 'black_bread', 'buckwheat', 'bun', 'cabbage_c', 'cabbage_salad', 'chicken_cutlet', 'compote', 'egg_with_mayonnaise_and_peas', 'fried_wings_-2_pieces-', 'halibut_with_lemon', 'herring', 'homemade_cutlet', 'jelly', 'lemon_drink', 'liver_in_an_omelet', 'mashed_potatoes', 'olivier_salad', 'orange_juice', 'pasta', 'pickle', 'pork_schnitzel', 'pork_with_mayonnaise', 'potato_soup', 'potatoes_with_meat', 'salad_with_radishes_and_herbs', 'salad_with_salmon_and_crackers', 'sausage_in_the_dough', 'steam_meatballs-2_pieces-', 'tea', 'tomato_and_cucumber_salad', 'tomato_juice', 'tomato_salad_with_cheese', 'vinaigrette', 'white_bread')

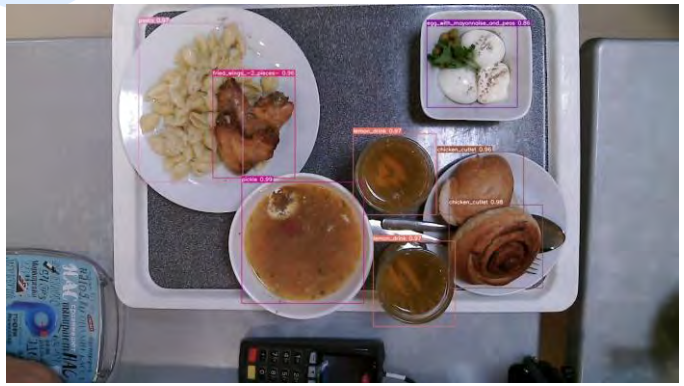
If there are a lot of object classes, and there is not enough data for training, then no architecture can cope. Need to use a custom classifier.

Yolo v5s



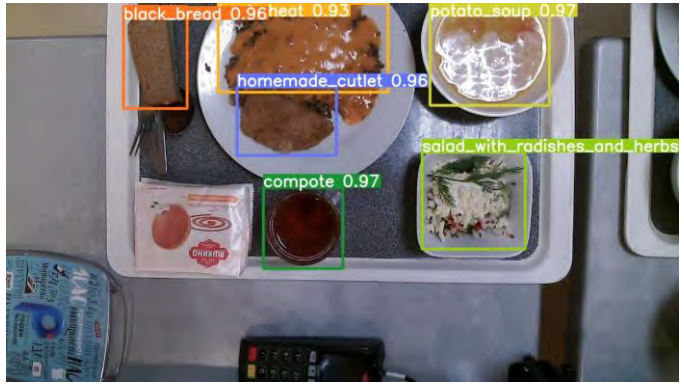
Recall: 0.856
Precision: 0.723
Map: 0.906

Yolo v6s



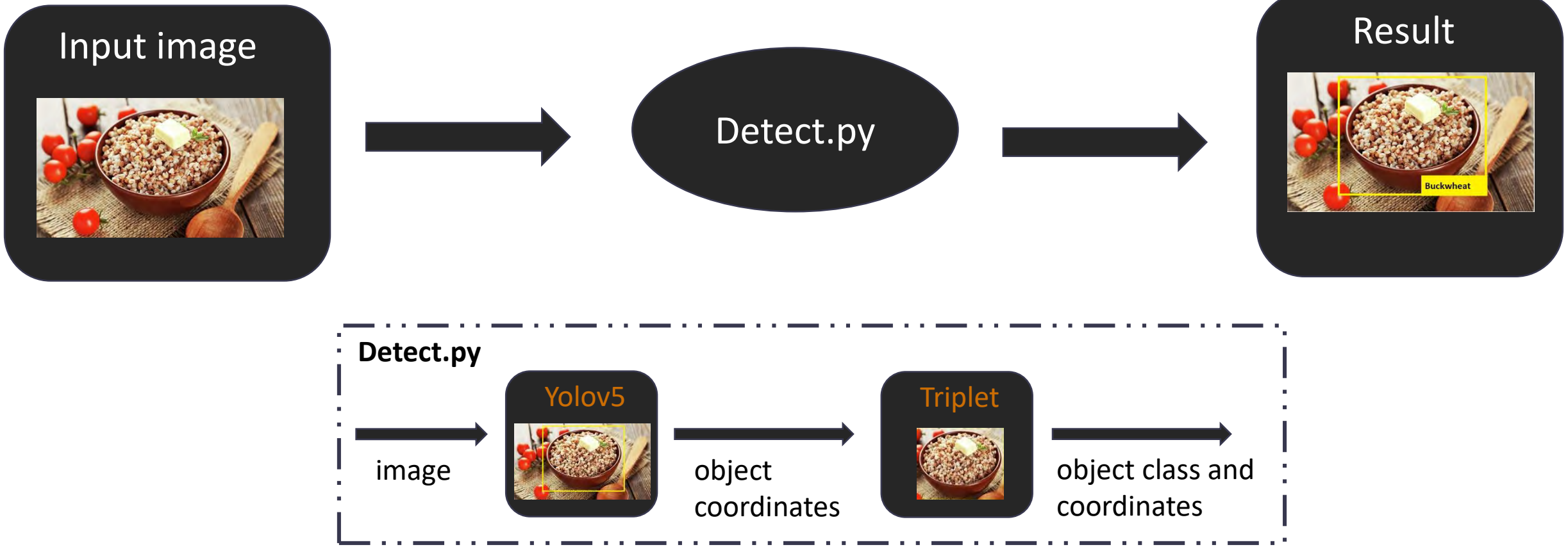
Recall: 0.805
Precision: 0.687
Map: 0.854

Yolo v7



Recall: 0.901
Precision: 0.845
Map: 0.925

The principle of operation



Final results

Yolo v5s

Metrics results

Recall	0.856
Precision	0.723
mAP	0.906

Yolo v6s

Metrics results

Recall	0.805
Precision	0.687
mAP	0.854

Yolo v7

Metrics results

Recall	0.901
Precision	0.845
mAP	0.925

Yolo + Triplet

Metrics results

Recall	0.995
Precision	0.994

Examples

Input images

Processing

Output images

Контроль загрязнения тяжелыми металлами

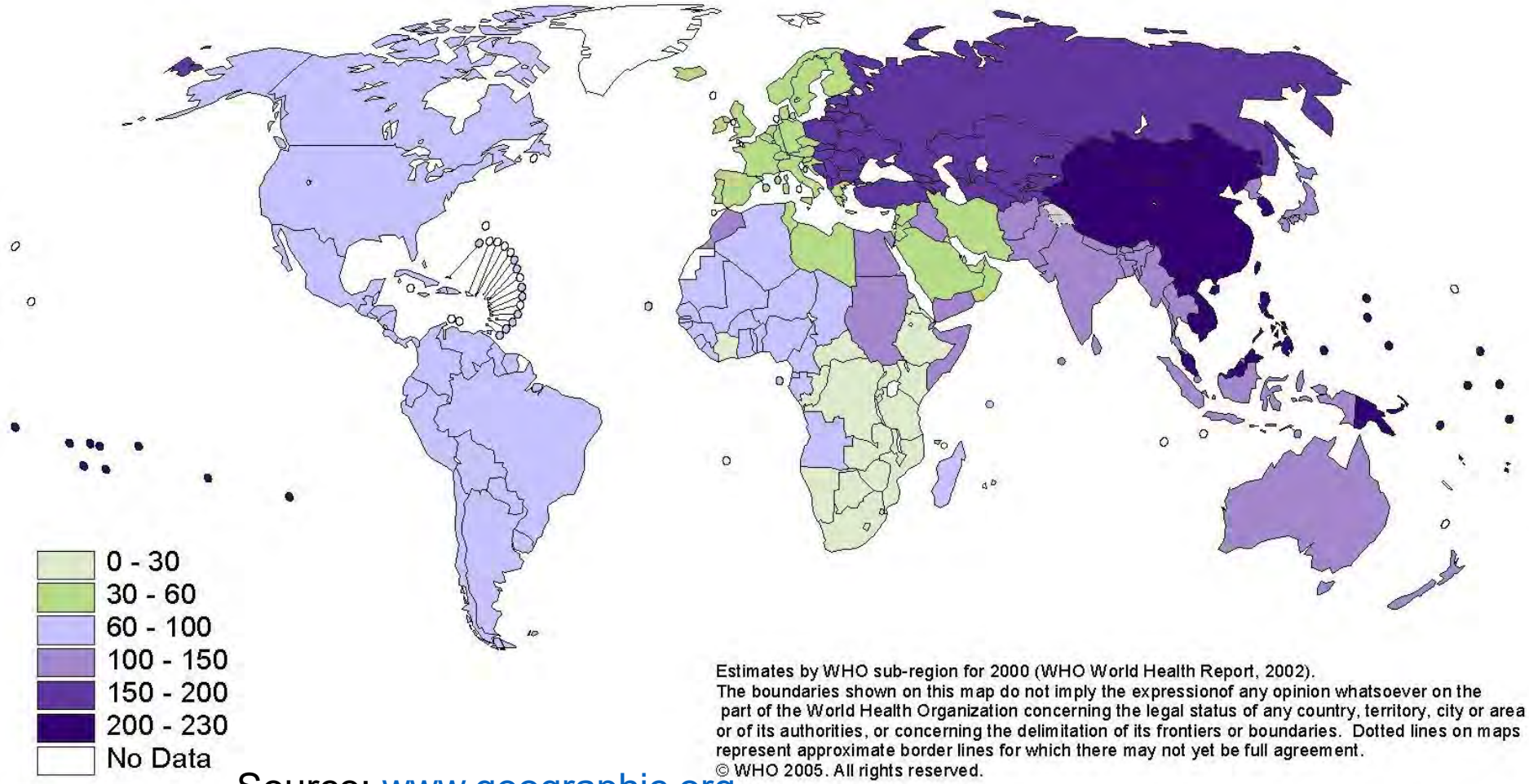
Introduction

Air pollution has a significant **negative impact** on the various components of ecosystems, **human health**, and ultimately, cause significant **economic damage**.

More than nine out of 10 of the world's population – 92% – lives in places where **air pollution exceeds safe limits**, according to research from the World Health Organization (WHO).

There are a lot of regional and international **environment control programs**. They use different techniques and tools but as a result, they all want to understand **what is the current situation** and how it will evolve.

Deaths from urban Air Pollution



Source: www.geographic.org

Approaches

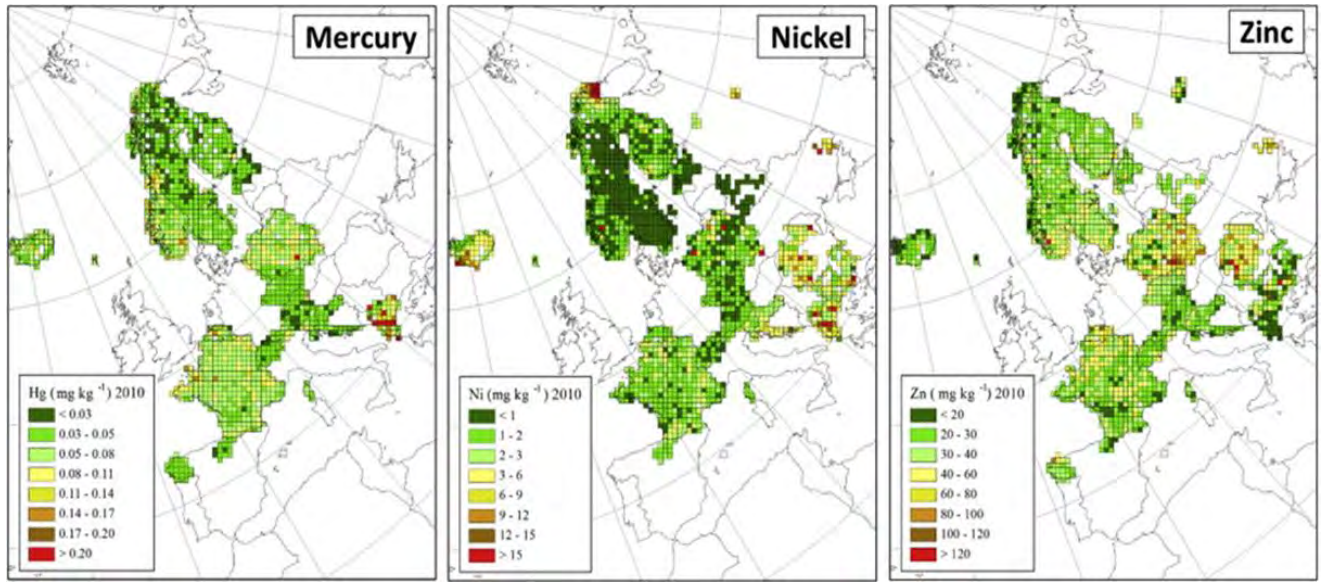
Generally, studies are based on the data obtained at the sampling sites in manual or automatic mode. The collected material is analyzed using various techniques in the field or in special laboratories. Air quality (AQ) monitoring stations provide information about regulatory air pollutants such as gaseous pollutants, PMs, and rarely about heavy metals. To get detailed information samples should be processed in laboratories.

After collection, the data are aggregated and interpreted, and quite often the results are ambiguous and require the involvement of experts.

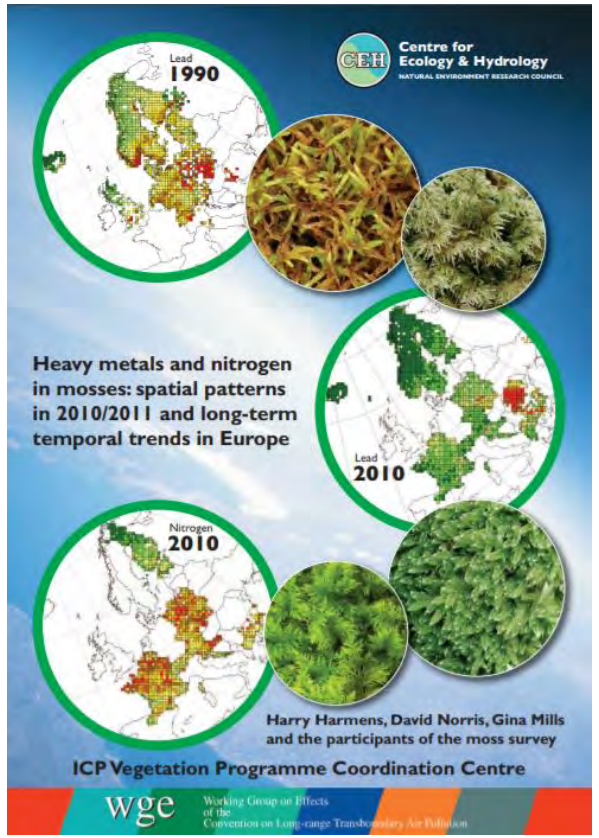
In most cases such kind of researches are limited, both spatially and temporally

ICP Vegetation

The aim of the **UNECE International Cooperative Program (ICP)** Vegetation in the framework of the United Nations Convention on Long-Range Transboundary Air Pollution is to **identify the main polluted areas of Europe**, produce regional maps and further develop the understanding of the long-range transboundary pollution. Atmospheric deposition study of heavy metals, nitrogen, persistent organic compounds (POPs) and radionuclides is based on the analysis of naturally growing mosses through moss surveys carried out **every 5 years**. The program is realized in **39 countries of Europe and Asia**. Mosses are collected at thousands of sites

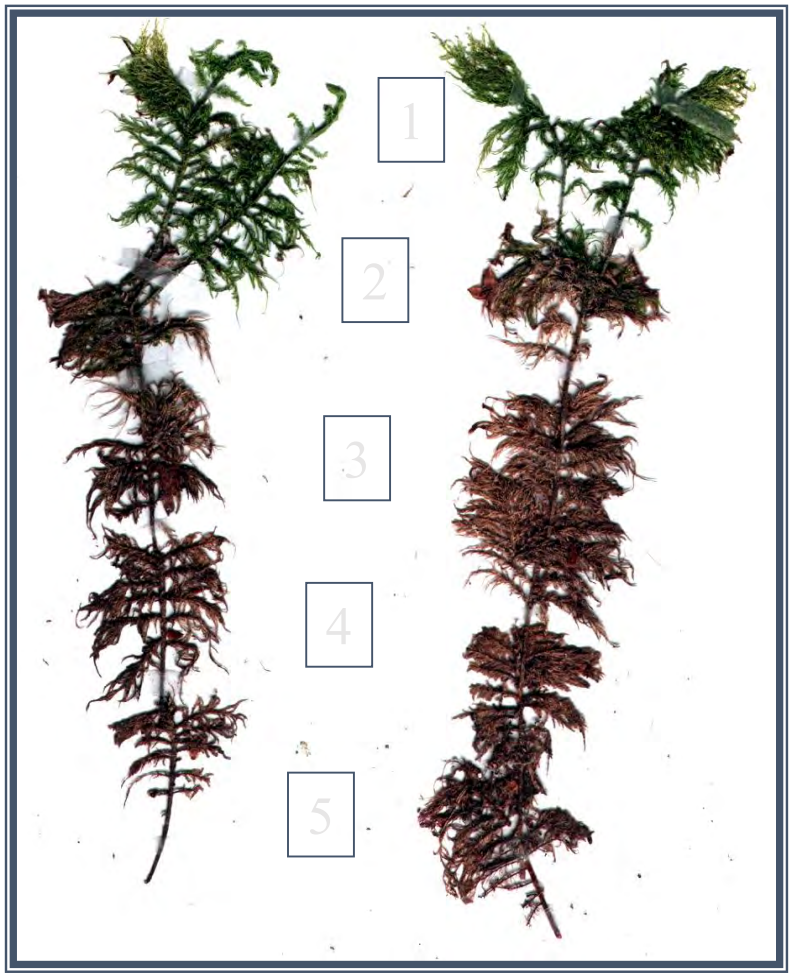


Examples of distribution maps in old Atlas



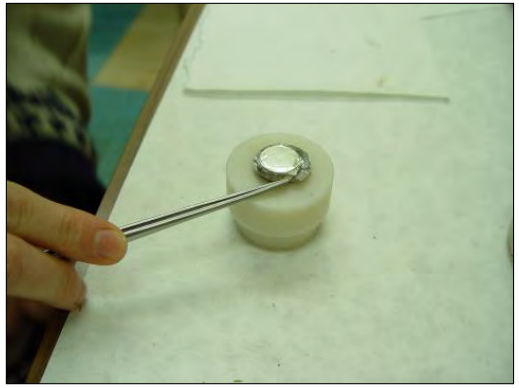
Since 2014 the JINR Frank Laboratory of Neutron Physics sector of neutron activation analysis is the **coordinator of the ICP** Vegetation program

Moss biomonitor

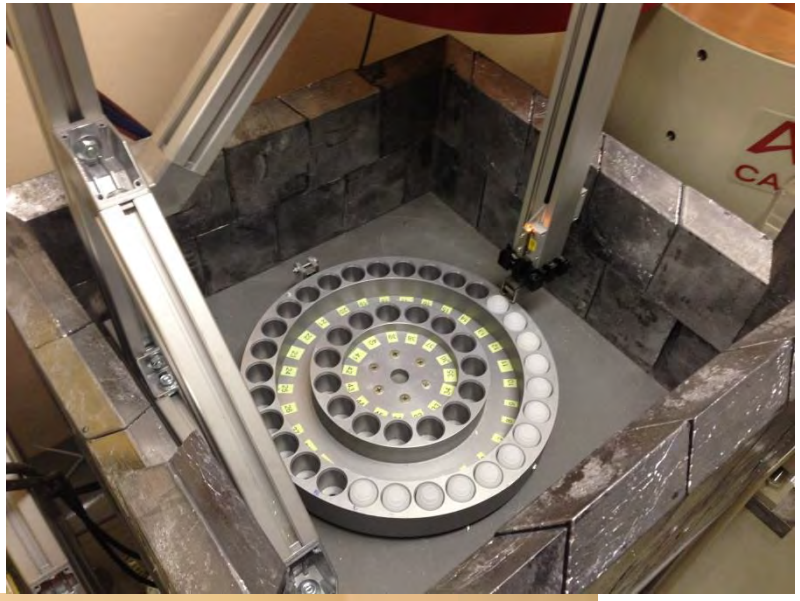


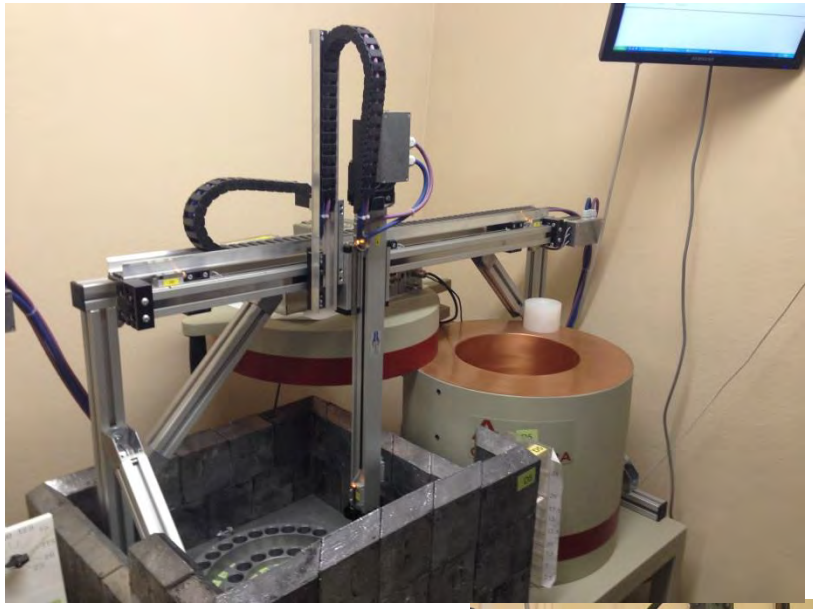
Annual segments

Sampling




1.
1.
3



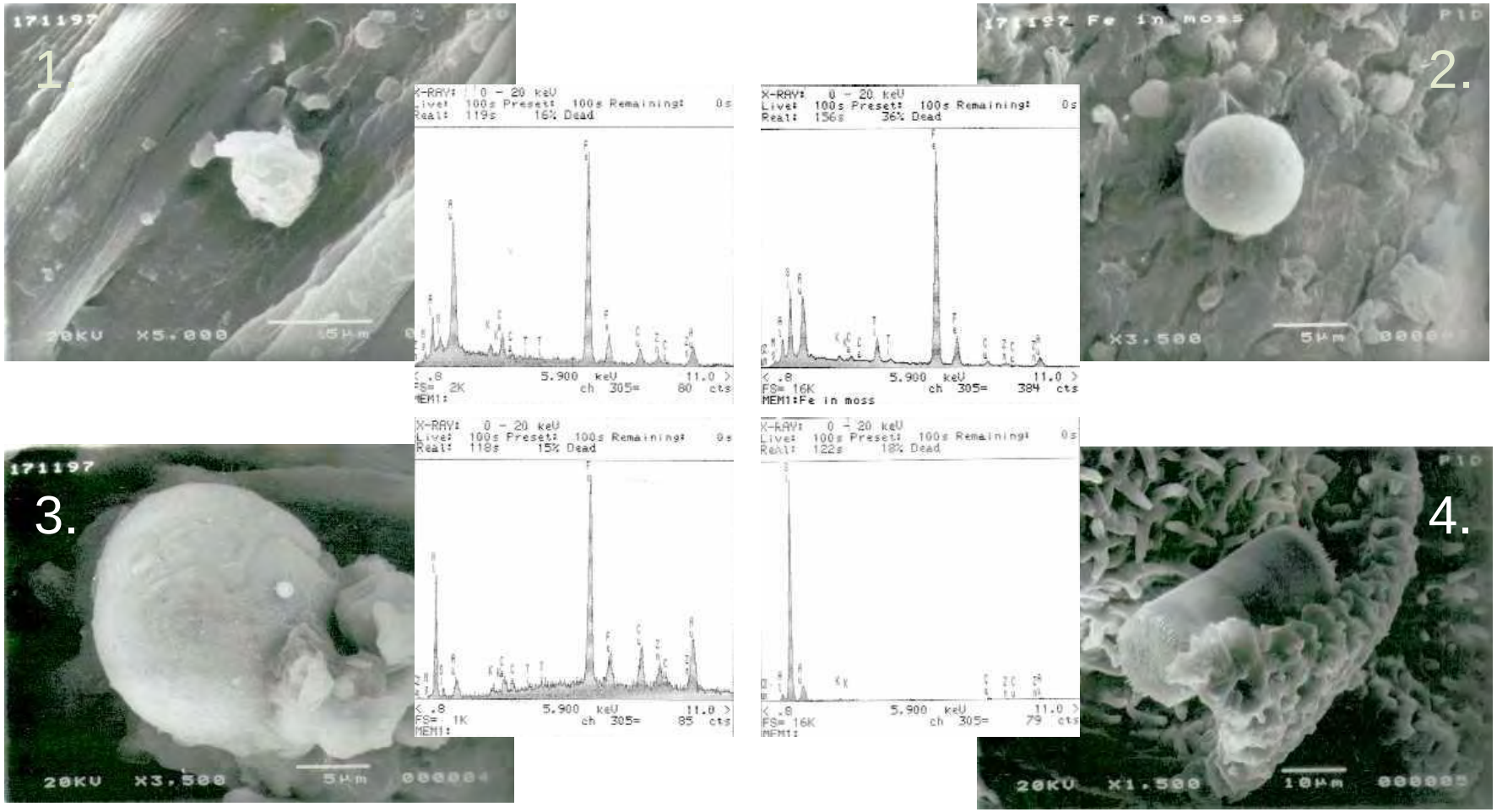


Three sample changers were installed

Each sample changer consists of:

- two axes liner movement device M202A (DriveSet, Germany)
- Rotated disk with 40 cells for samples (JINR)
- Three axes Xemo Motion controller with software and cables (Systec GmbH, Germany)

Scanning electron microscope images of captured particles on the moss surface and corresponding spectrograms



1 - Fe particle with Mg impurity; **2** - Spherule of pure iron;
3 - Al-Fe cluster particle with impurities of Zn, Cu, and Ti; **4** - Diatomic alga

ICP Vegetation (Past)

The UNECE ICP Vegetation program had a serious drawback related to its **weak adoption of modern informational technologies**. Information on collecting and processing of samples was carried out **manually** or with minimum automation.

Until 2016, data mostly was stored in Excel files. It was aggregated and processed in different packages (ArcGIS, MATLAB, etc.) **manually by the coordinator**.

Files from respondents were usually passed to the coordinator **by email**. There were **no common standards** in data transfer, storing and processing software.

Such a situation does not meeat the modern standards for quality, effectiveness, and speed of research and demands developing a **single platform** to provides a comprehensive solution for biological monitoring and forecasting tasks

ICP Vegetation (First steps)

The idea was to create a cloud platform for data management to facilitate IT aspects of all biological monitoring stages starting from a choice of collection places and finishing with generation of pollution maps of a particular area or state-of-environment forecast in the long term

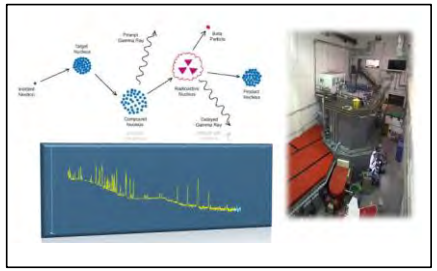
We was going to move from simple tasks:
- storing and manipulating with data,
- processing of data,
- calculation of basic statistics,
- creation of simple maps
- etc.

to more complicated one:
- optimization of the sample collection spatial distribution
- advanced mathematical methods for multi-level intelligent statistical analysis,
- geostatistical analyses,
- atlases and reports creation
- and others.

The platform now

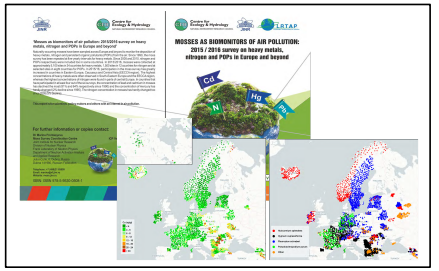
Samples collection

Samples analysis

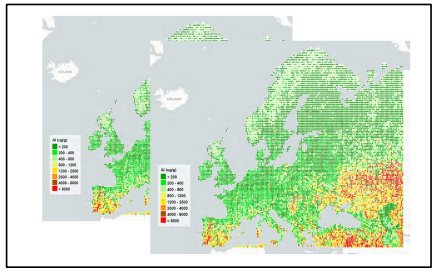


Data analysis

Data presentation

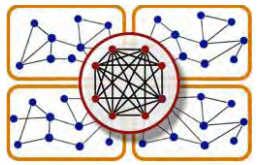


Prediction/Controle



API HTTP

php AJAX python™



Tensorflow
Keras

Microservices

mongoDB.

NGINX

Open Nebula

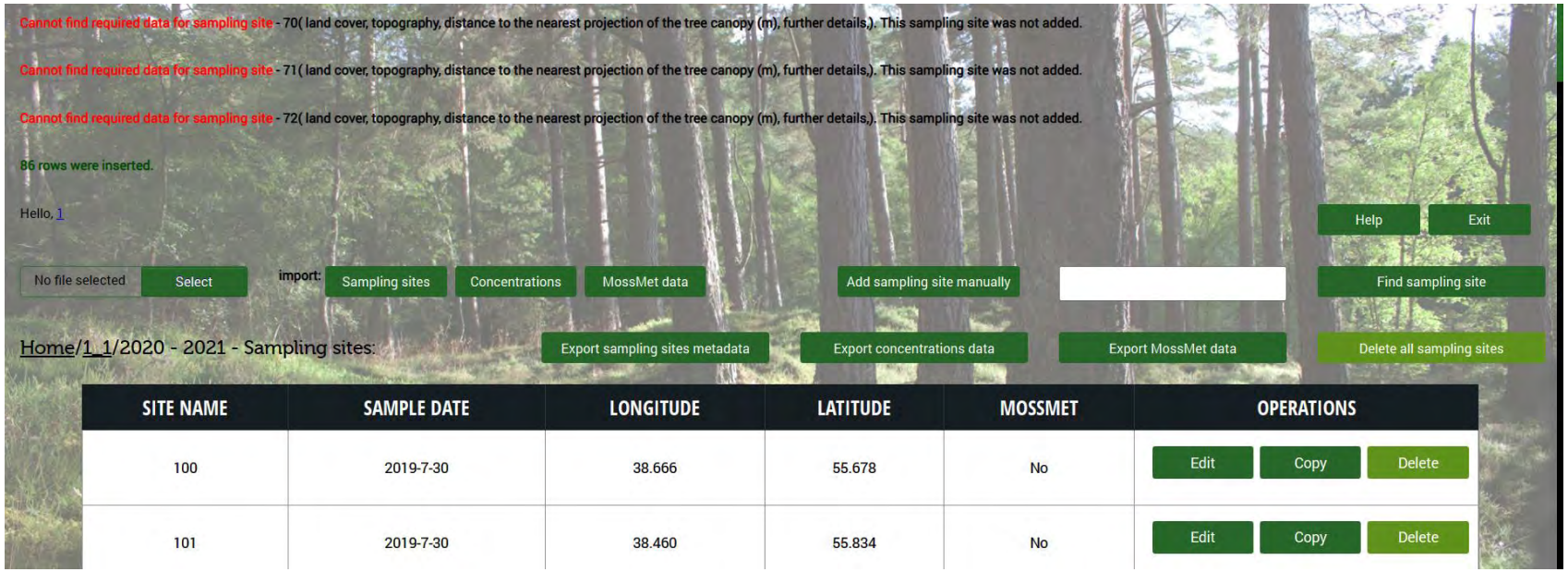
Google Earth Engine

HybriLIT

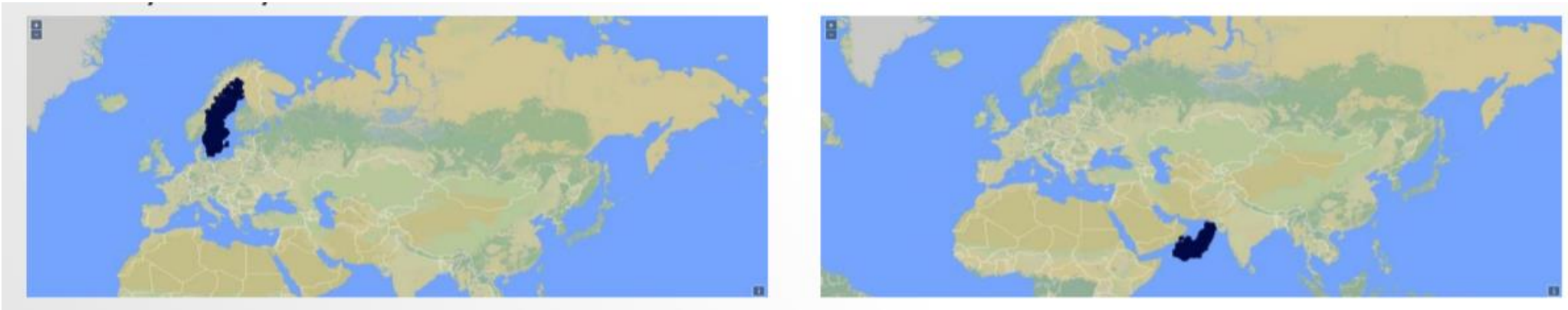
Since the launch of the first version of the platform, a mobile application has been developed to simplify the process of collecting and verifying data, deep learning models for image classification and pollution prediction based on remote sensing data, various functional blocks implemented in a microservice architecture to automate a number of operational tasks, and the analytical capabilities of the system are also expanded.

WHY IT'S USEFULL

Fast verification of data structure and it completeness



Easy way to find human made mistakes

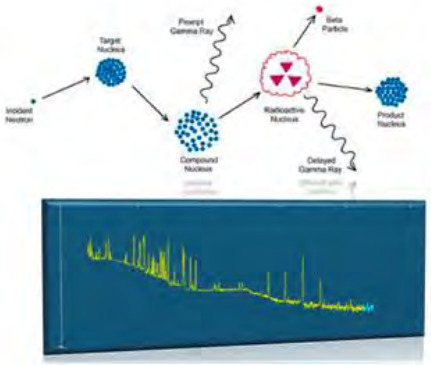


- Nice tool to analyze data. • Access to yours data from anywhere. • Online processing and results.
- Ability to store historical data and analyze trends

Workflow

1. Collecting samples

2. Processing samples

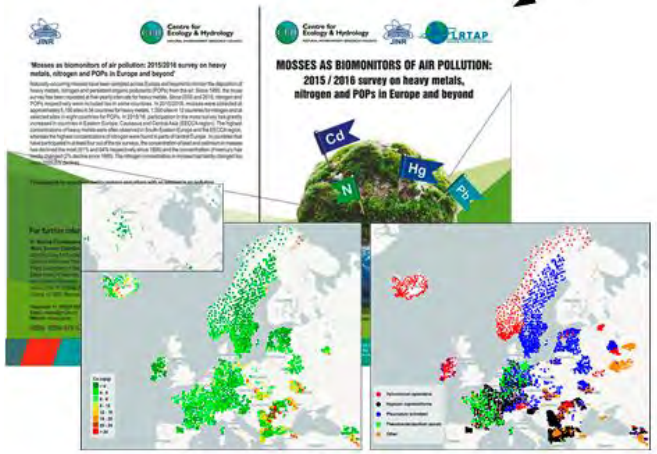


Samples metadata

HM concentrations

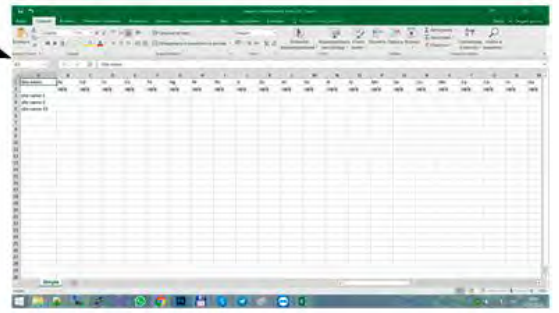
moss.jinr.ru

5. Reports



3. Filling in the information about the concentrations

Site names

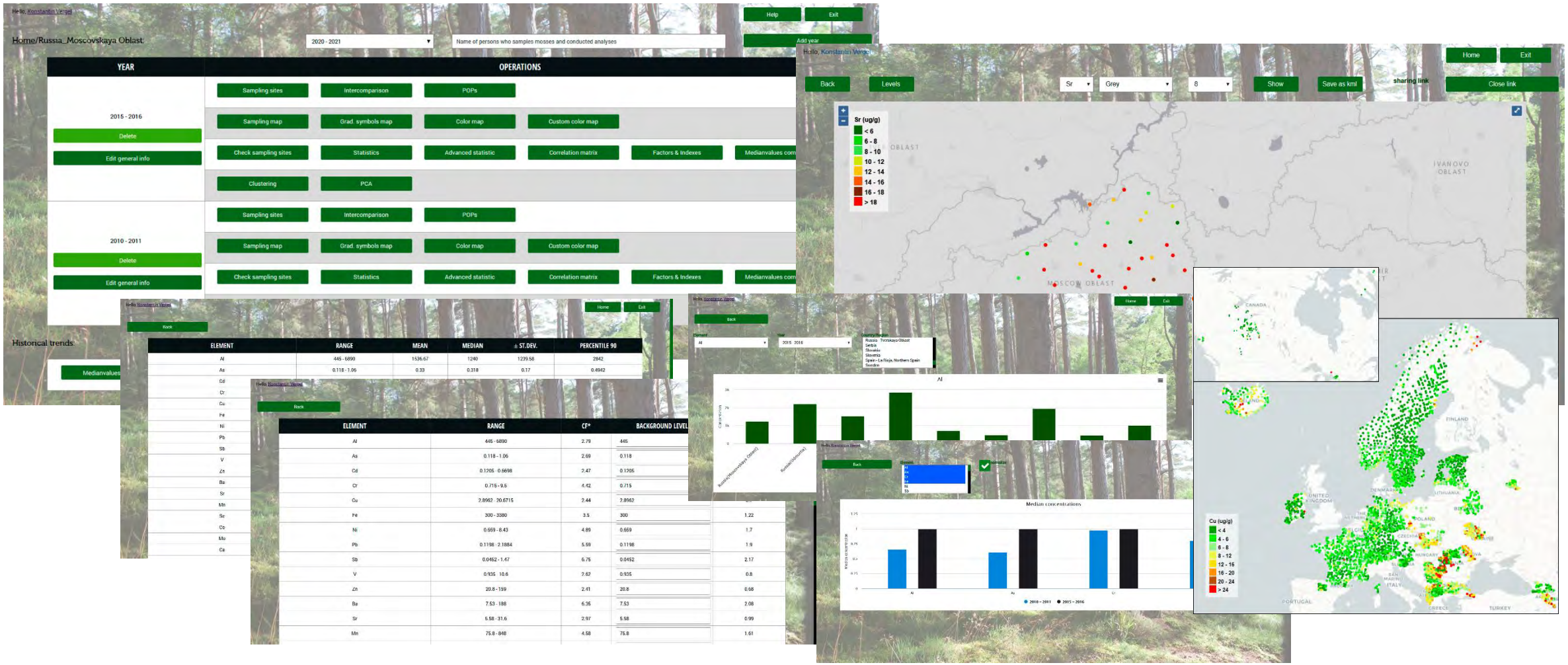


HM concentrations

4. Analysis

DMS

The Data Management System (DMS) of the UNECE ICP Vegetation was developed at the Laboratory of Information Technologies and consists of a set of interconnected services and tools deployed and hosted at the Joint Institute for Nuclear Research (JINR) cloud infrastructure. DMS is intended to provide its participants with a **modern unified system of collecting, analyzing and processing of biological monitoring data.**



DMS. Atlas 2015-2016

JINR

CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

'Mosses as biomonitors of air pollution: 2015/2016 survey on heavy metals, nitrogen and POPs in Europe and beyond'

Naturally-occurring mosses have been sampled across Europe and beyond to monitor the deposition of heavy metals, nitrogen and persistent organic pollutants (POPs) from the air. Since 1990, the moss survey has been repeated at five-yearly intervals for heavy metals. Since 2005 and 2010, nitrogen and POPs respectively were included too in some countries. In 2015/2016, mosses were collected at approximately 5,100 sites in 34 countries for heavy metals, 1,500 sites in 12 countries for nitrogen and at selected sites in eight countries for POPs. In 2015/16, participation in the moss survey has greatly increased in countries in Eastern Europe, Caucasus and Central Asia (EECCA region). The highest concentrations of heavy metals were often observed in South-Eastern Europe and the EECCA region, whereas the highest concentrations of nitrogen were found in parts of central Europe. In countries that have participated in at least four out of the six surveys, the concentration of lead and cadmium in mosses has declined the most (81% and 64% respectively since 1990) and the concentration of mercury has hardly changed (2% decline since 1995). The nitrogen concentration in mosses has hardly changed too since 2005 (5% decline).

This report is for scientists, policy makers and others with an interest in air pollution.

For further information or copies contact:

Dr Marina Frontasyeva
Moss Survey Coordination Centre
Joint Institute for Nuclear Research
Division of Nuclear Physics
Frank Laboratory of Neutron Physics
Department of Neutron Activation Analysis and Applied Research
Joliot-Curie, 6, Dubna, Russia
Dubna 141980, Russian Federation

Telephone: +7 (496)21 65609
Email: marina@nf.jinr.ru
Website: moss.jinr.ru

Dr Harry Harmens
ICP Vegetation Coordination Centre
Centre for Ecology & Hydrology
Environment Centre Wales
Deiniol Road
Bangor
Gwynedd LL57 2UW
United Kingdom

Telephone: +44 (0) 1248 374500
Email: hh@ceh.ac.uk
Website: icpvegetation.ceh.ac.uk

ISBN: ISBN 978-5-9530-0508-1

CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

JINR

LRTAP
Long-range Transboundary Air Pollution

MOSSES AS BIOMONITORS OF AIR POLLUTION:
2015 / 2016 survey on heavy metals, nitrogen and POPs in Europe and beyond

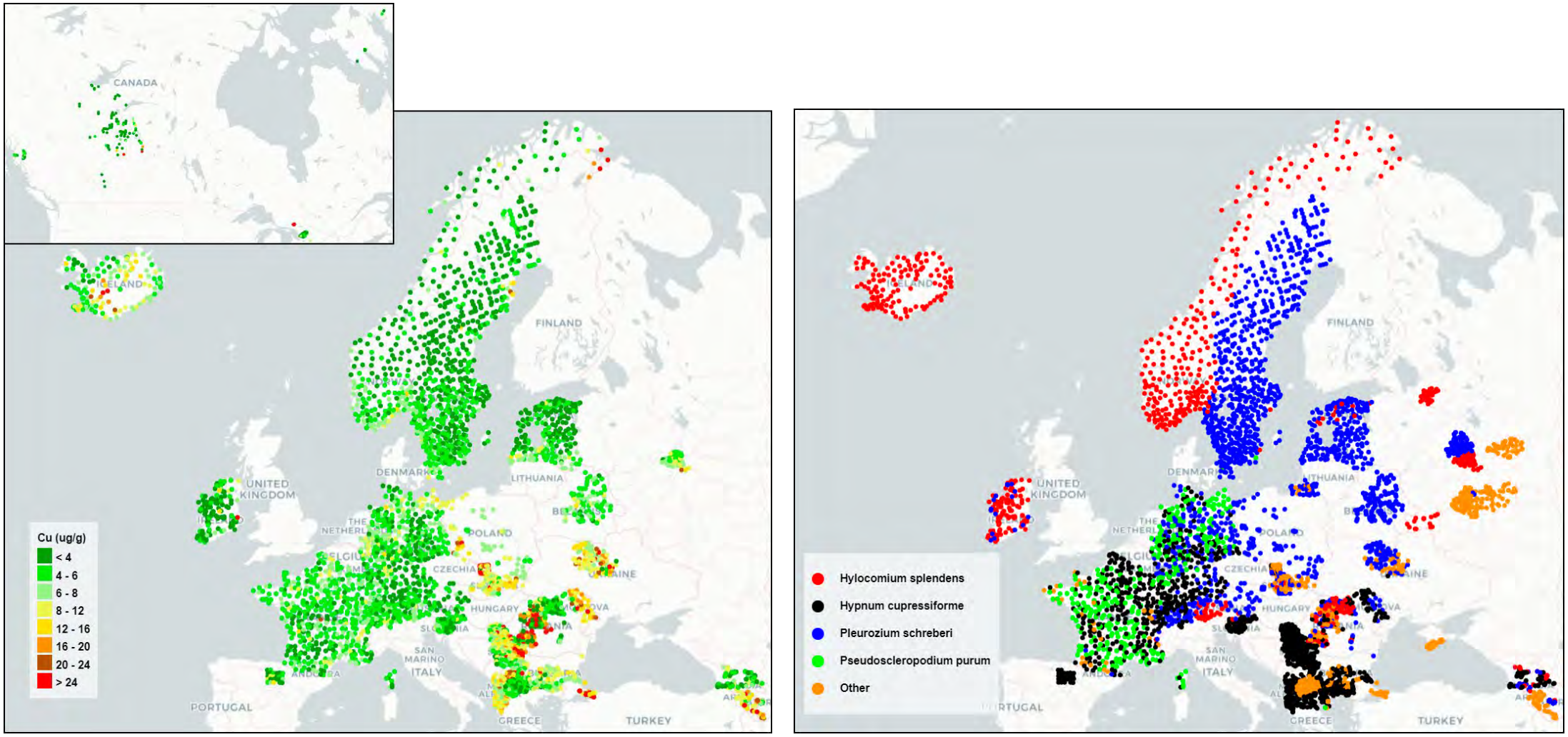
Cd
Hg
N
Pb

Marina Frontasyeva, Harry Harmens, Alexander Uzhinskiy and the participants of the moss survey

ICP VEGETATION

wge Working Group on Effects of the Convention on Long-range Transboundary Air Pollution

DMS. Maps



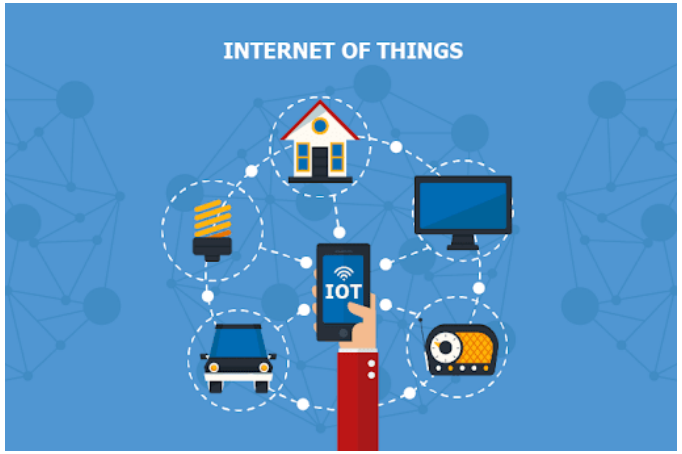
Examples of the maps for the Atlas 2015-2016

Flashback to Ind 4.0

Smart technologies

In last decade, various modern technologies are used in environmental pollution control projects, which make it possible to provide a new level of service, as well as the quality and speed of obtaining results. Now we can talk about intelligent platforms capable of generating new knowledge based on incoming and available data and, in some cases, making decisions that previously required the competence of an expert.

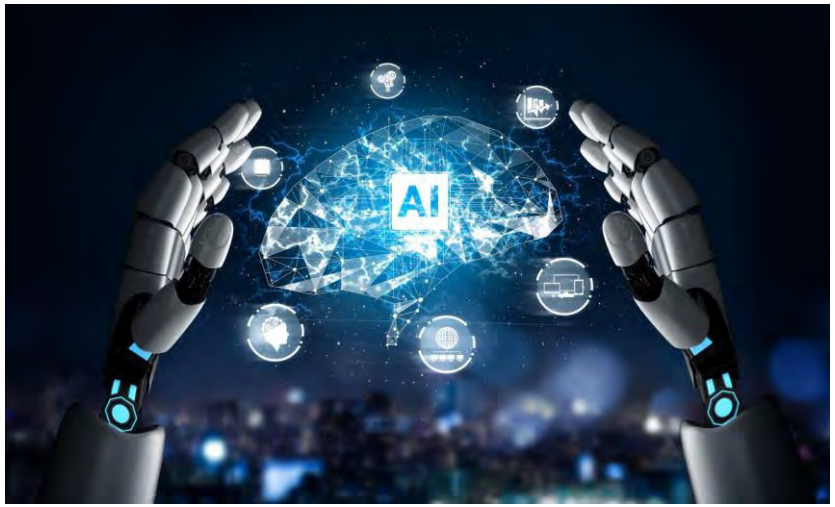
Here are only few examples of such technologies



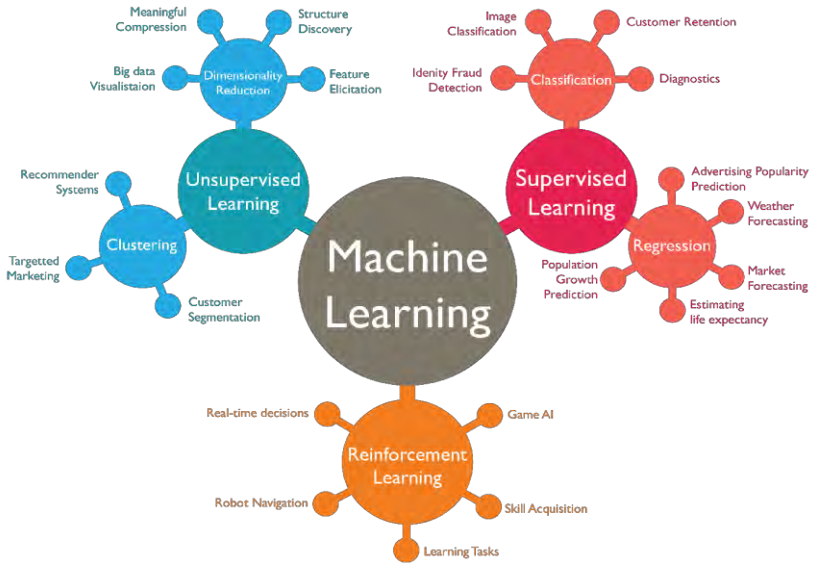
The Internet of things (IoT) specify the principles of connection and exchanging data between physical objects that are embedded with sensors and another objects, programs and systems. Many platforms use IoT technologies to organize sensor network and process environmental monitoring data. That allow to minimize number of errors, automate routine processes, and speed up data gathering routines.

The Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. In case of the environmental monitoring the data we have to work with could be both large if we dial with huge sensor network and complex if we dial with sampling sites meta-data.

Smart technologies



Artificial intelligence (AI) is a wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. In environmental monitoring there are always operations requiring expert opinion. AI technologies could execute primely analysis and save expert time.

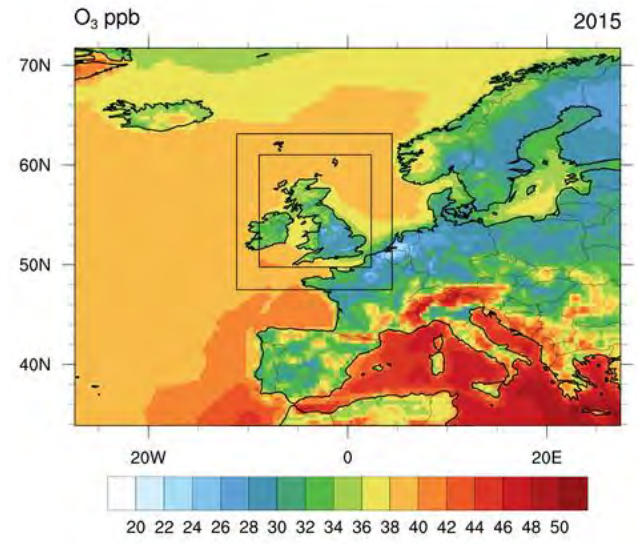


Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Both classification and prediction tasks of ML are very useful for environmental monitoring.

Robotics, remote sensing, drones etc.

Smart technologies

Here are few examples of such platforms:



MegaFon offers a platform for environmental monitoring based on the Internet of Things - MegaFon.Ecology.

SimpliVity promotes Omnicube, a universal smart monitoring solution that allows to effectively control various aspects of enterprise operations, including environmental parameters.

Rostec is implementing projects in the field of intelligent environmental monitoring systems.

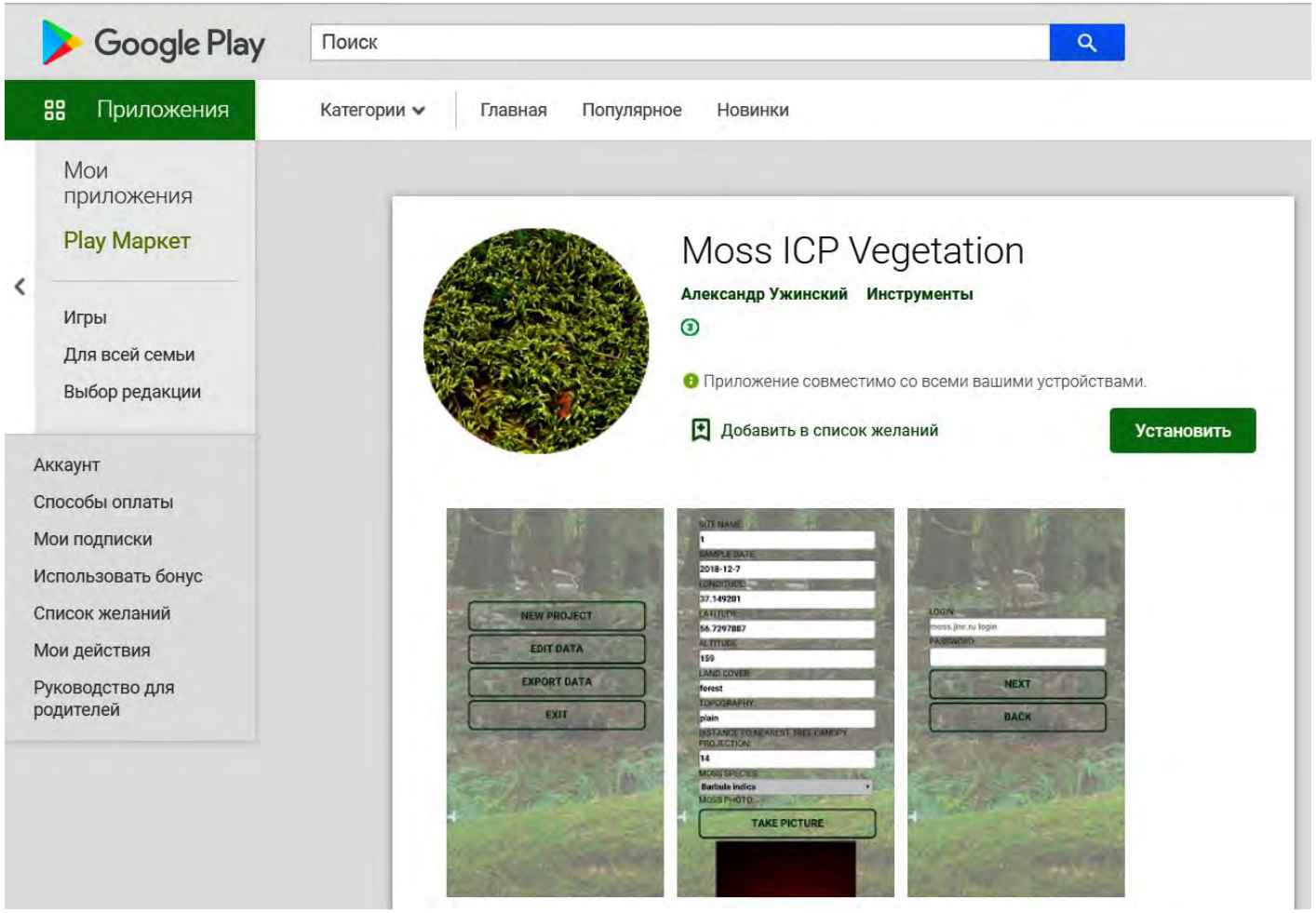
There are solutions that combine weather stations of various levels into a single infrastructure.

Naturally, there are also foreign projects, mainly based on IoT technologies. For example, the platform of EXM and Libelium companies, designed to improve the efficiency of environmental monitoring, or solutions from the Filippetti Group or Novolyze, providing similar functionality.

EMEP (*European Monitoring and Evaluation* Programme) use transport models and Air control station data for atmospheric transport and deposition modeling.

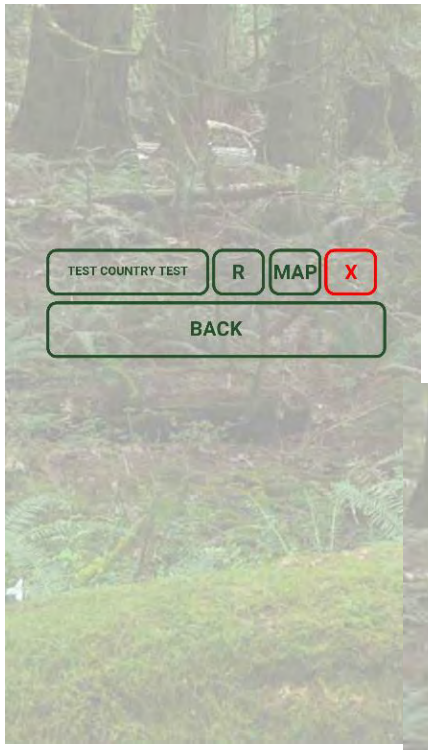
Mobile application

During the processing of data for Atlas 2015-2016, we experienced misspelling of moss names, wrong coordinates, negative concentrations, and many other problems with data. Now for Atlas 2020-2021, we have the mobile application that allows filling in as required by the UNECE ICP Vegetation manual information about sampling sites.



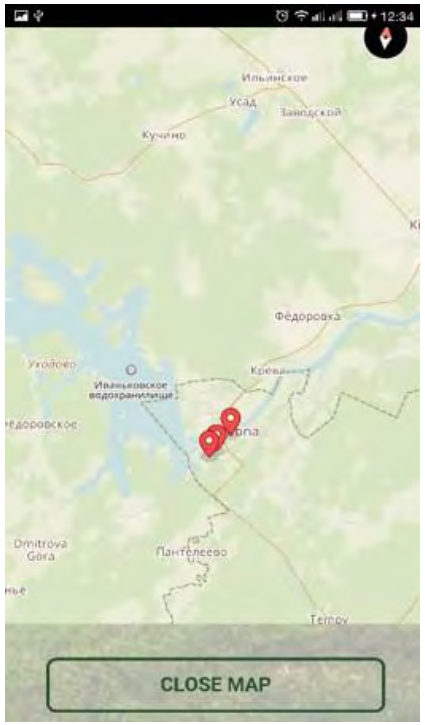
The application automatically sets longitude and latitude of the sampling site, controls the correctness of the input data and allows capturing photos of the moss samples and the nearest area. The application integrated with the DMS and all information about sampling sites can be imported to the system.

Mobile application



Create/edit project

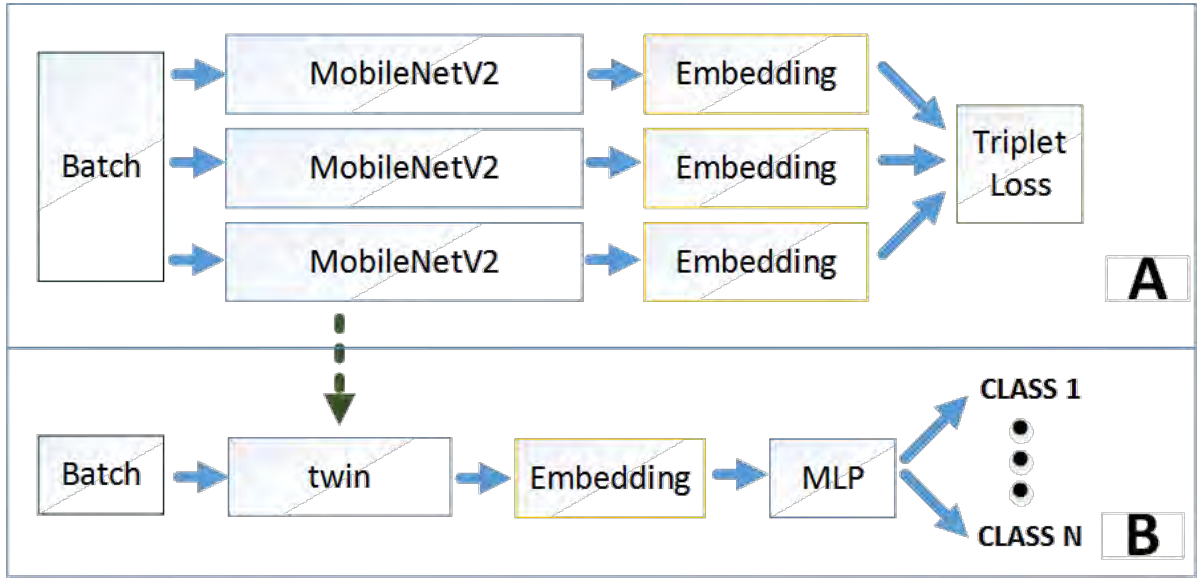
- Site name should be unique
- MossMet information could be added



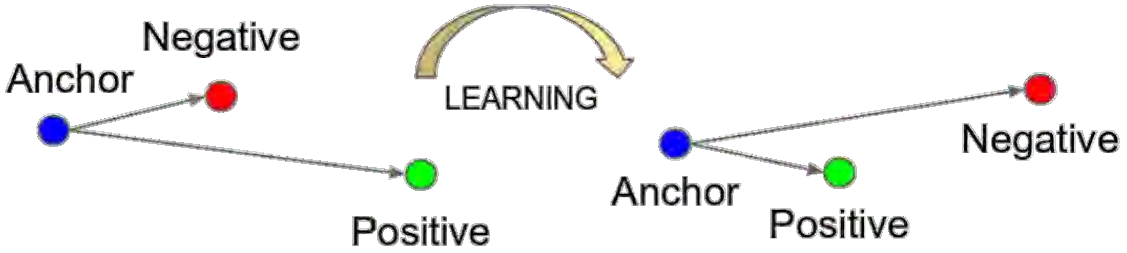
Create/edit sampling sites

Interesting tasks (Moss species classification)

- 599 images
- 5 moss species
- **97.6% classification accuracy**



Siamese networks with triplet loss function



$$L = \max(d(a,p) - d(a,n) + margin, 0)$$

"d" is some kind of function for calculating the distance between vectors, for example, Euclidean distance.

A **"a"** is an anchor image which we want to identify

P **"p"** image the same class as anchor

N **"n"** image of another class not matching the anchor

Interesting tasks (Storing of complex data + geodata)


```
{
  "_id": ObjectId("60a124f83c549b478a4b4d9e"),
  "user_id": ObjectId("58981bdf9e7ba441018b4dca"),
  "project_id": "58981c239e7ba443018b53ab",
  "year_id": "5be5331f9e7ba476718b4926",
  "site name": "31. Aiviekste",
  "longitude": 25.9442,
  "latitude": 56.6528,
  "sample date": "2020-09-22",
  "altitude (m)": 83,
  "land cover": "Forests-coniferous",
  "topography": "plain",
  "distance to the nearest projection of the tree canopy (m)": 3,
  "further details": "sunny",
  "moss species": "Pleurozium schreberi",
  "cd": 0.123,
  "cr": 0.3991,
  "cu": 6.3528,
  "fe": 123.8097,
  "ni": 0.2276,
  "pb": 0.6013,
  "v": 0.3835,
  "zn": 24.7611
...
}
```



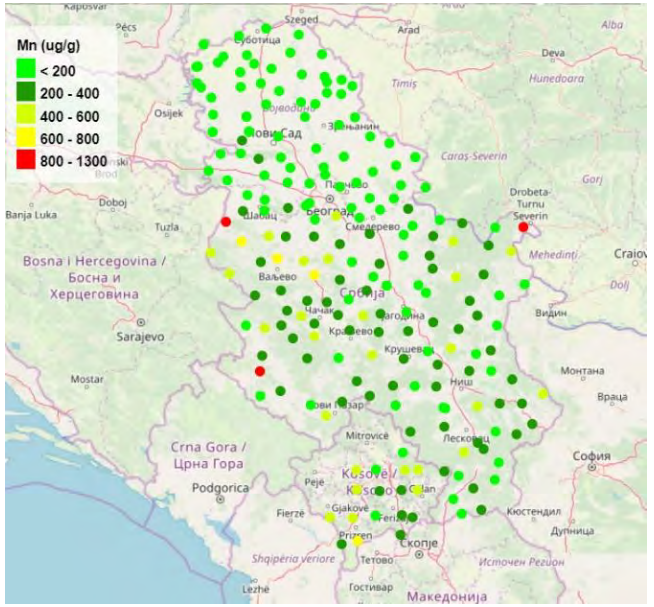
```
db.Address.find({
   "mapLocation": {
     $geoWithin: {
       $box: {
         [ [bottom left coordinates], [top right coordinates] ]

       }
     }
   }
})
```

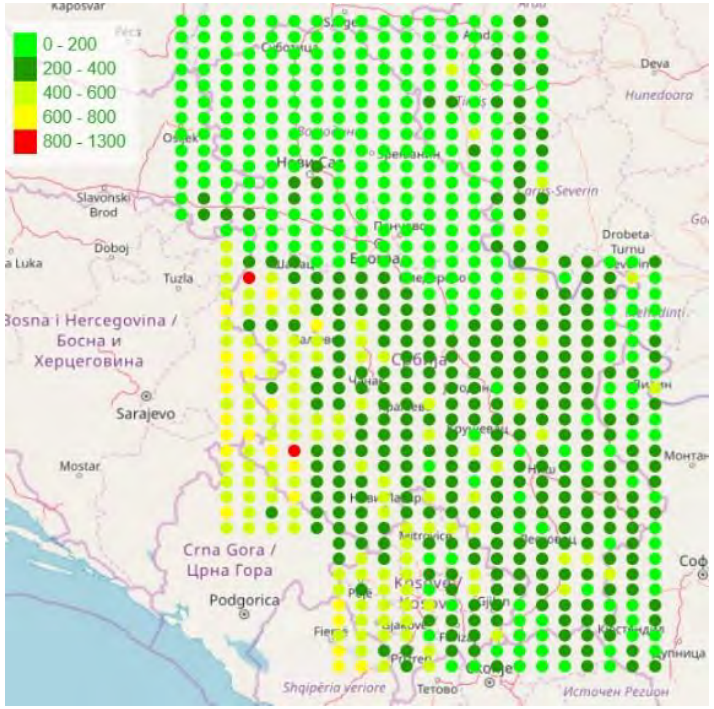
- Different collections (sampling sites, PoP's, Intercomparison etc)
- Tens to hungered parameters
- Geo-spatial data
- Satellite imagery indexes – hungered thousands to millions objects

Interesting tasks (prediction)

Regulatory monitoring of air pollution by potentially toxic elements (PTEs) are limited, both spatially and temporally.

Modelling of air pollution can be a good option for overcoming gaps in the data gathering, while moss bag biomonitoring has been recognized as a technique for highly spatially resolved measurements of PTE air pollution.



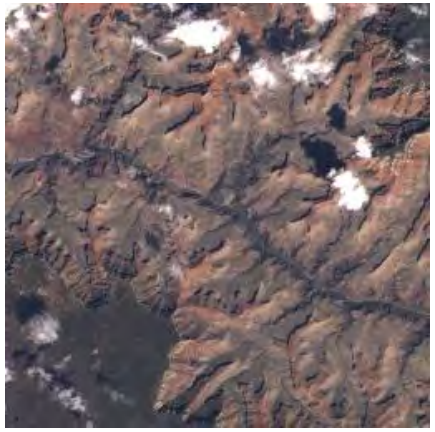


Modelling allows us to:

- monitor the evaluation of situation when it needed,
- get detailed information about areas of interests,
- check the situation at the cross border areas,
- partly automate the environment control process.

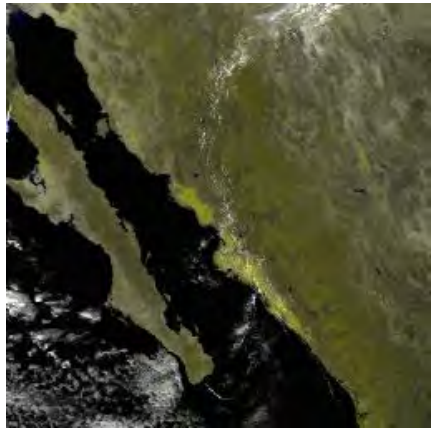
Google Earth Engine

There are more than **100 satellite programs** and modeled datasets. Google Earth Engine has **JavaScript online editor** to create and verify code and **python API** to communicate with user's applications.

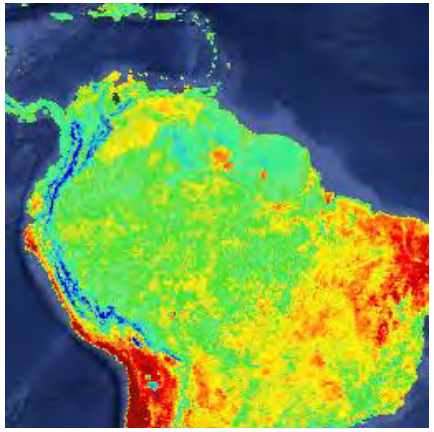


Landsat (15-30m Resolution)

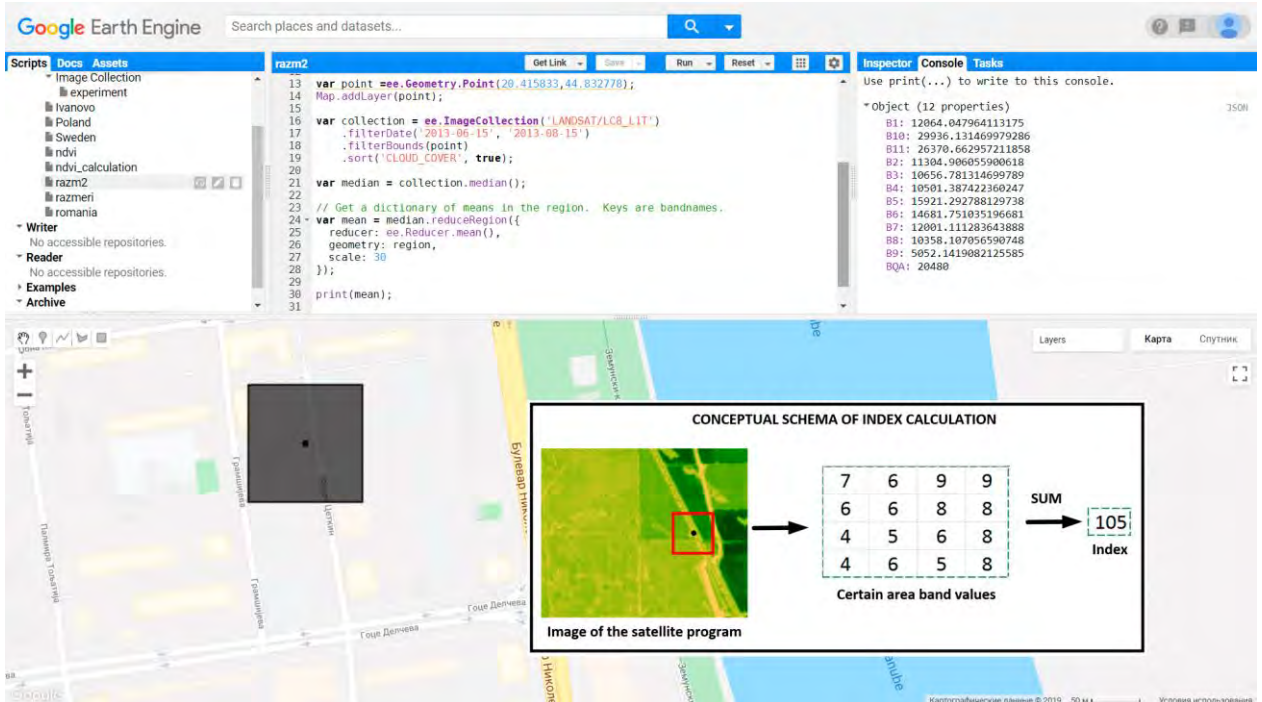
Modis (250-500m Resolution)



Sentinel (250-500m Resolution)



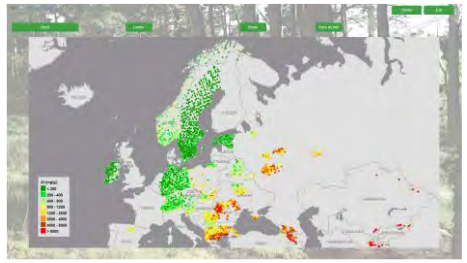
The MOD11A2 V6 average 8-day land surface temperature (LST) in a 1200 x 1200 kilometer grid.



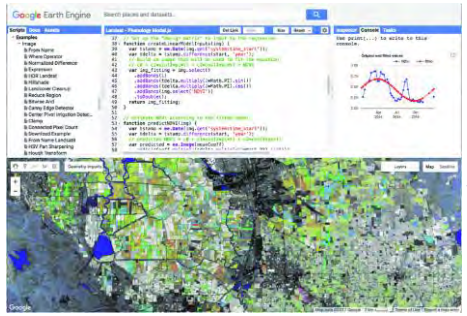
Specify program and time-period to get a collection of images, for example, program – "MODIS/006/MOD09A1" from 2013-06-15 to 2013-08-15 (the period relevant for in situ biomonitoring). Then, define the analyzed area, for example, a square kilometer, with center at the coordinates where sampling was performed. During the satellite data collection, under the bands (channels) of the median image, we execute some mathematical functions (max, min, median, etc.) and get the numerical values.

Schema

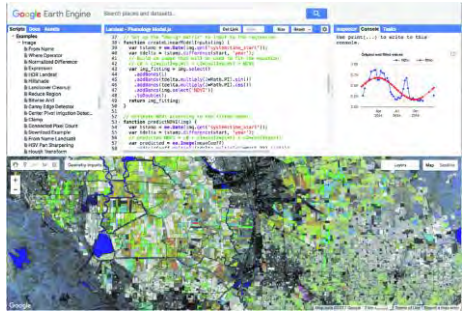
We use **satellite imagery data** and the **artificial neural network** to **predict concentration**. The general idea is to use data that we can get from satellite images together with sampling data from DMS to learn NN and then use only data from satellite images to predict concentration.



DMS



Google Earth Engine



Google Earth Engine

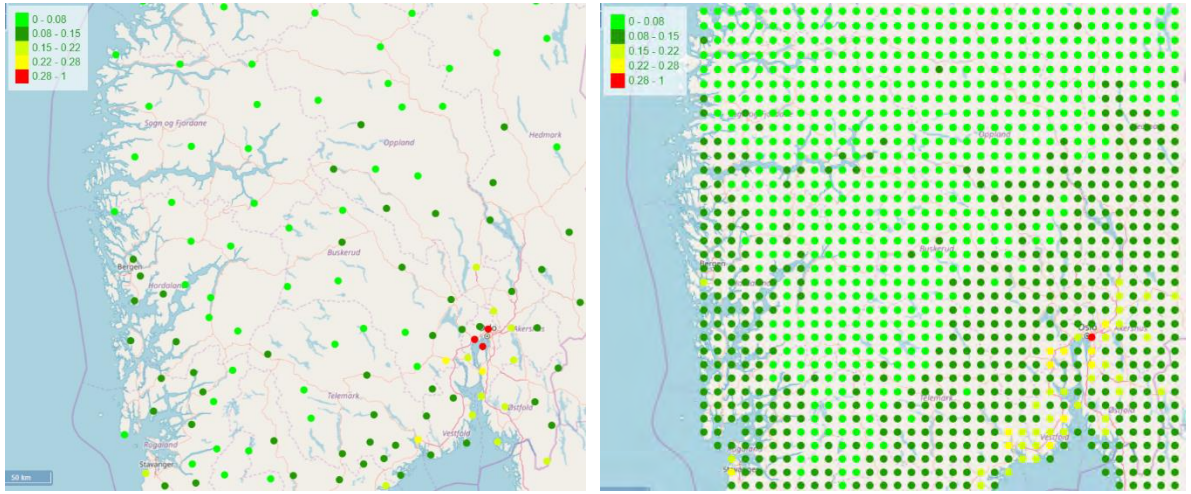
Training

Learned model

Learned model

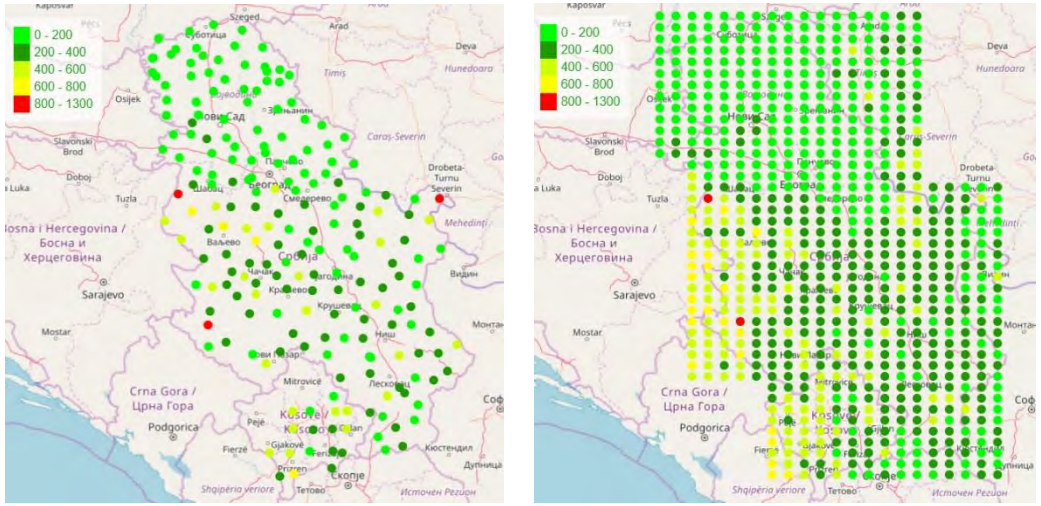
Prediction

Results on the regional level

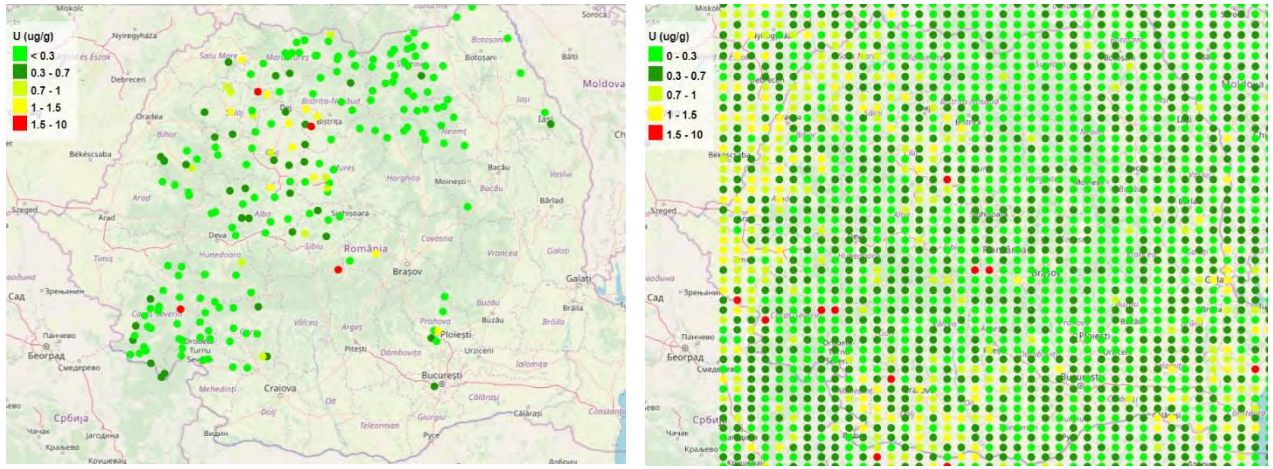


Sb at Norway. Left – real life, right - prediction

Candidates for modeling:
Al, As, Cr, Cu Fe, Mn, Ni, Pb, V, Sb, U …



Mn at Serbia. Left – real life, right - prediction



U at Romania. Left – real life, right - prediction

Urban Level (Belgrade)

The goal of this study was to facilitate the highly resolved mapping of the presence of potentially toxic elements in the air of an urban area, which is typically characterised by high and variable pollution. + to check whether model can keep appropriate accuracy during long time period.

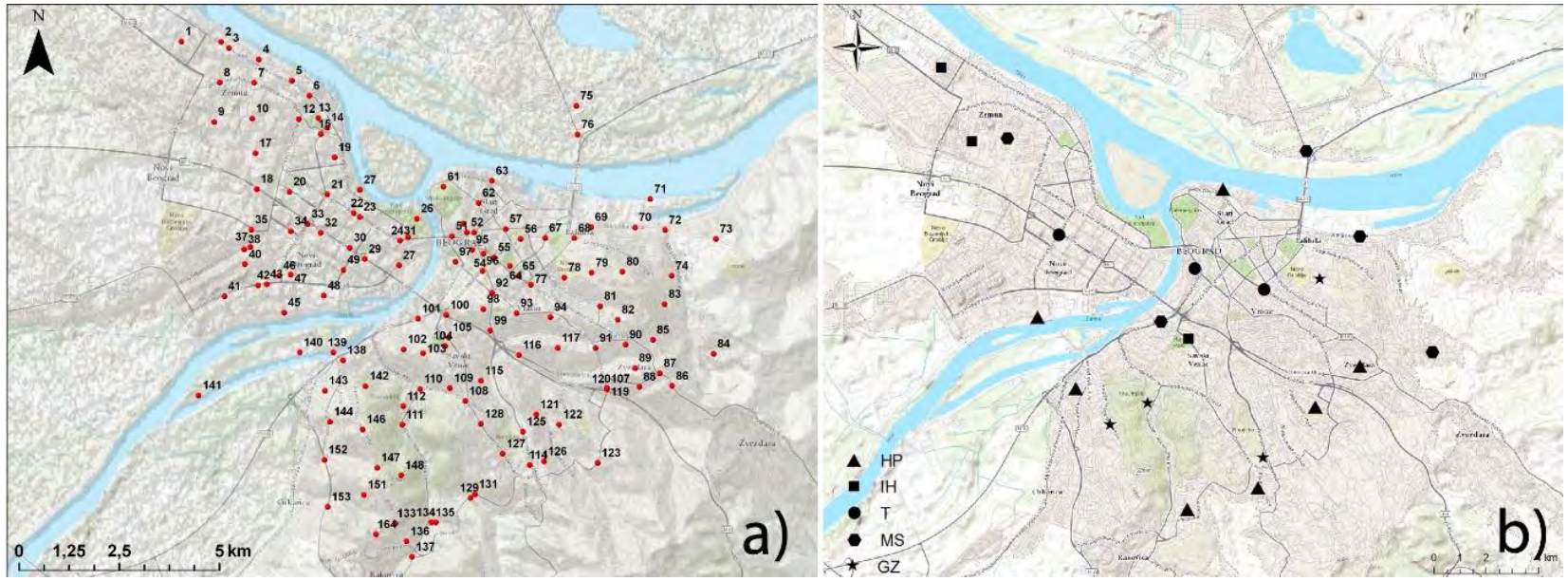


Figure 1. Moss bag biomonitoring across the Belgrade urban area; maps of the sampling sites during two seasons: (a) summer (urban, suburban and green zones) and (b) winter (U–urban sites, GZ–green zones)

Urban Level (Belgrade)

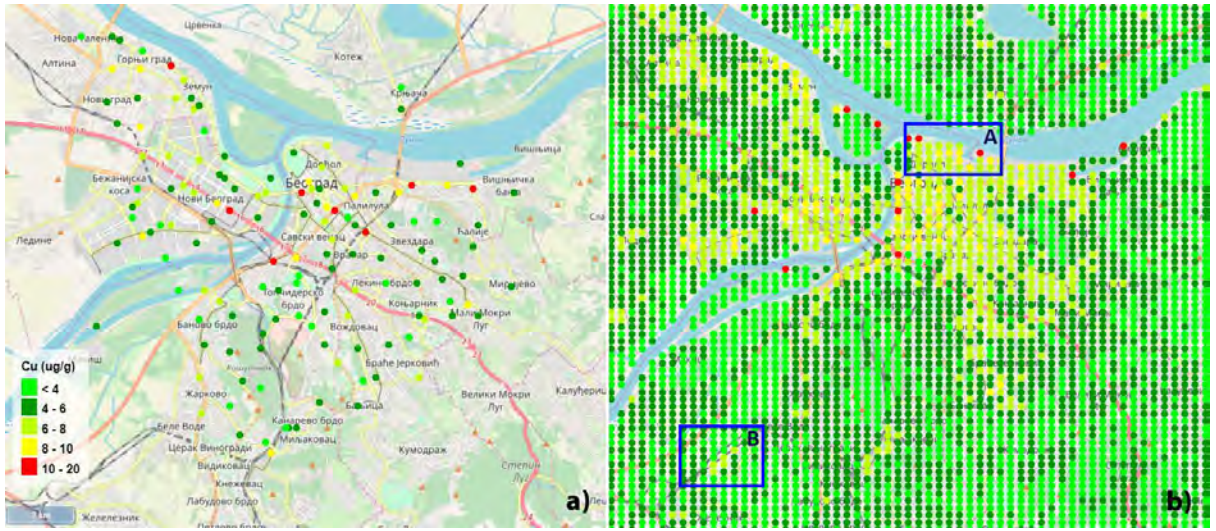


Figure 3. Concentration of Cu in the summer of 2013 (Belgrade): a) real measurements, and b) prediction values; area A represents central part of Old Belgrade with permanently high traffic flow; area B represents a large railway terminal

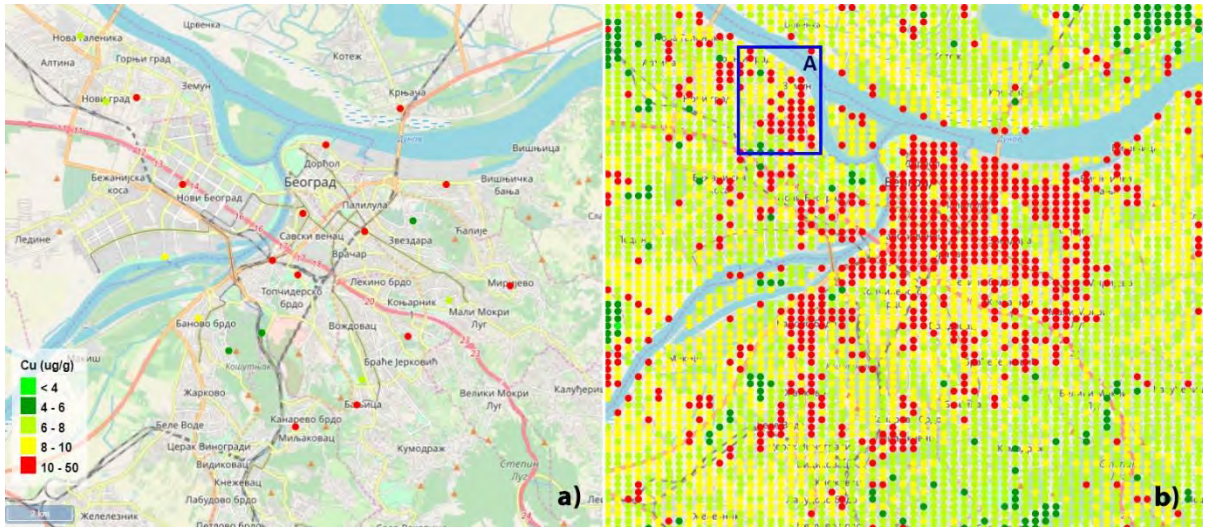


Figure 4. Concentration of Cu in the winter season 2013/2014 (Belgrade): a) real measurements, and b) prediction values; area A represents an old city core highly polluted in winter season

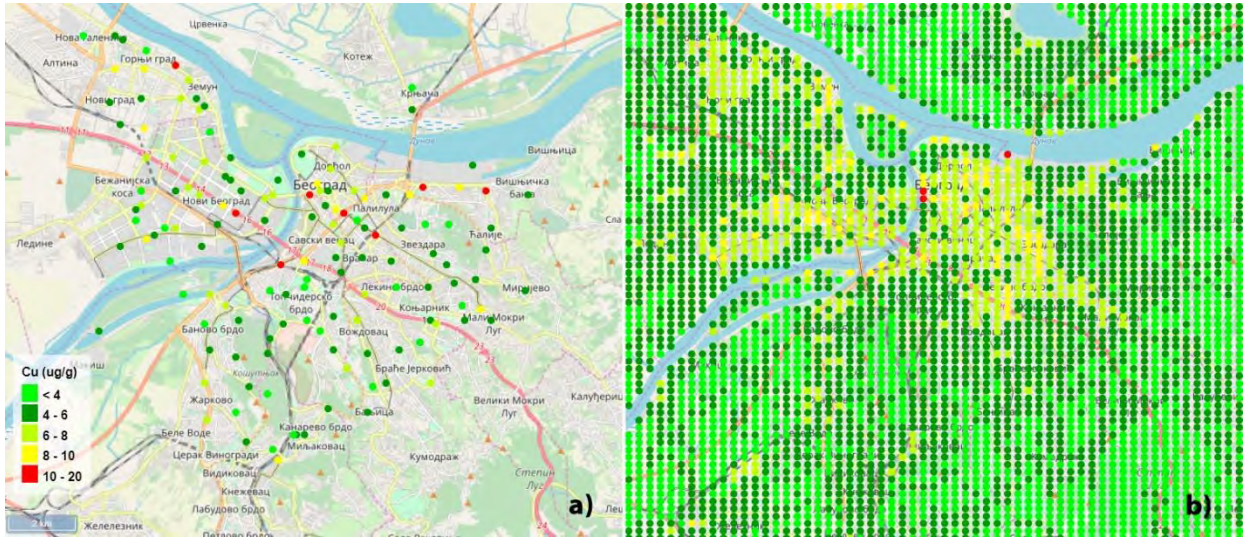
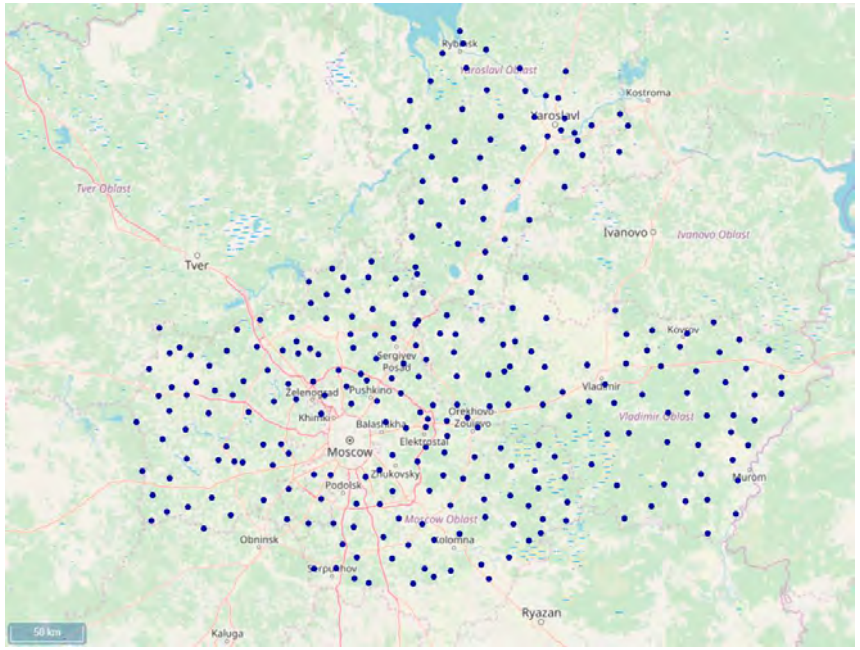


Figure 5. Concentration of Cu in Belgrade: a) biomonitoring measurements in the summer of 2013, and b) prediction for 2018

Machine learning and neural networks



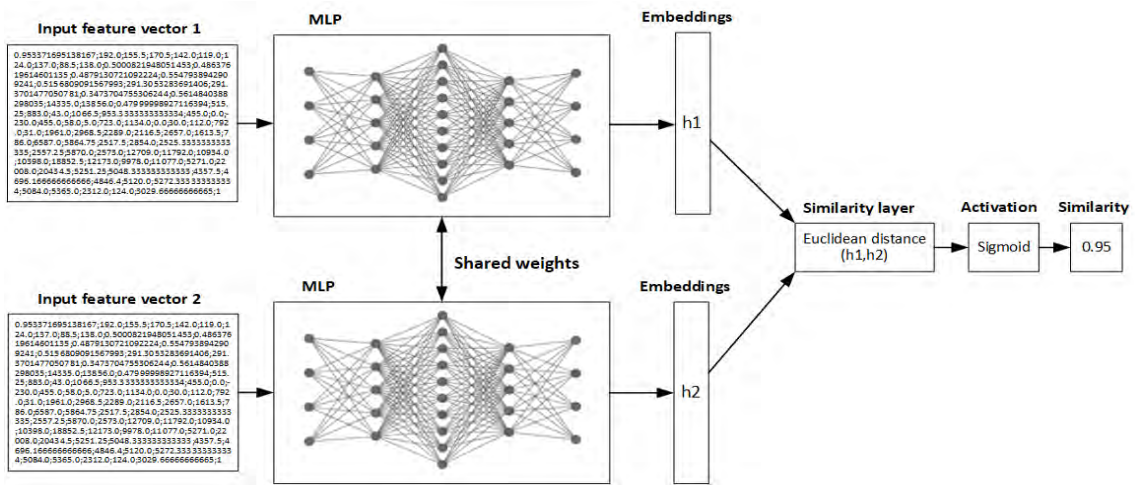
We use the information on 73, 53, and 156 samples from the Vladimir, Yaroslavl, and Moscow regions gathered in 2018 - 2019.

The indices are gathered based on data from 13 programs for 281 sampling sites, and their linkage with the concentration of 18 heavy metals is verified. Altogether 9 HMs, i.e., Al, Fe, Sb, Na, Sc, Sm, Tb, Th, and U, look very prospective for modeling.

We examine three approaches: Gradient Boosting, Multilayer perceptron, and Siamese network.

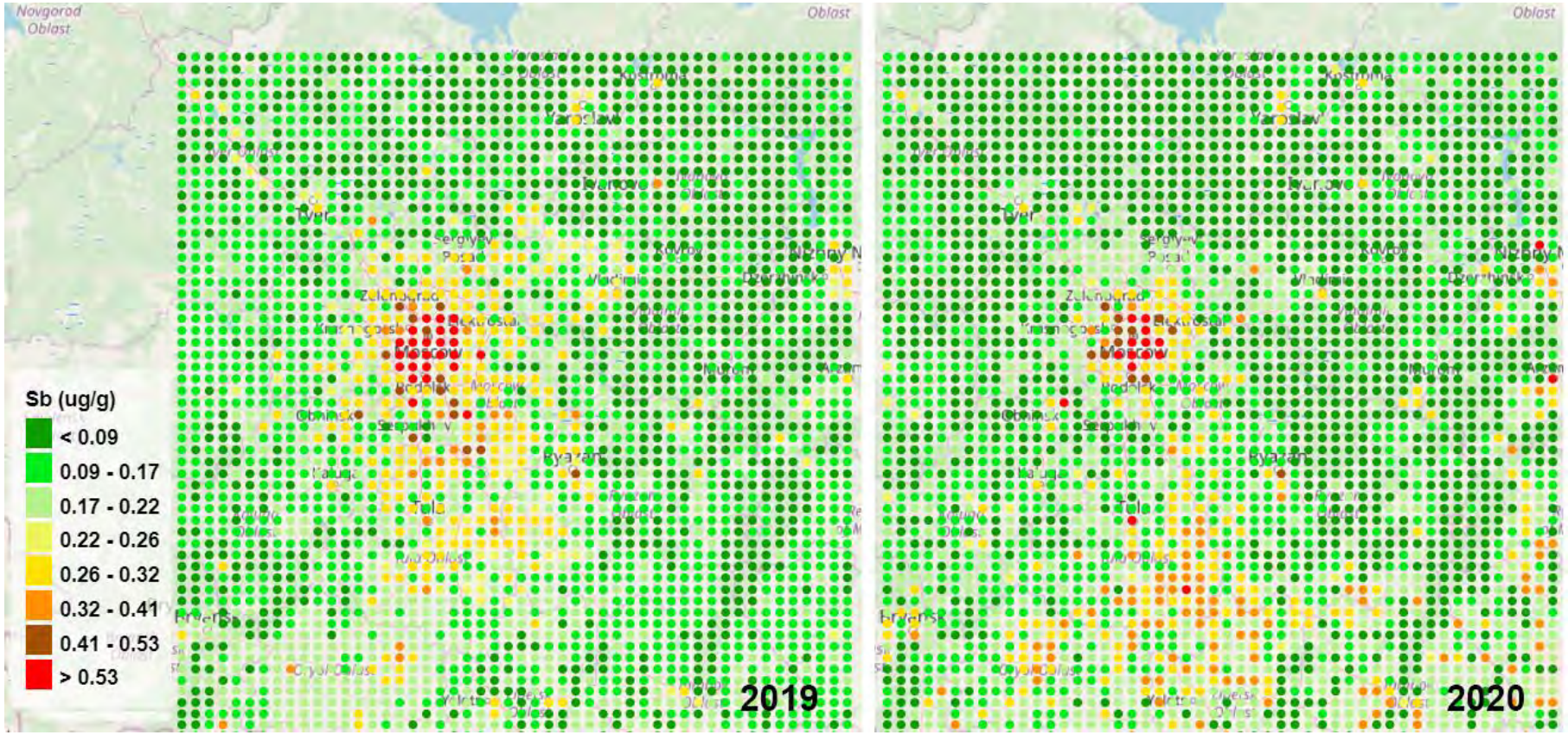
	Al		Fe		Sb	
	Acc si	Acc ai	Acc si	Acc ai	Acc si	Acc ai
GB	0.91	0.92	0.92	0.93	0.94	0.94
MLP	0.89	0.91	0.92	0.92	0.89	0.92
SNN	0.92	0.93	0.93	0.93	0.93	0.94

Table 2. Mean accuracy of the models. GB is gradient boosting. MLP is the multilayer perceptron. SNN is the Siamese neural network. Acc Si is the accuracy on the selected indices. Acc Al is the accuracy on all indices.



Siamese network architecture

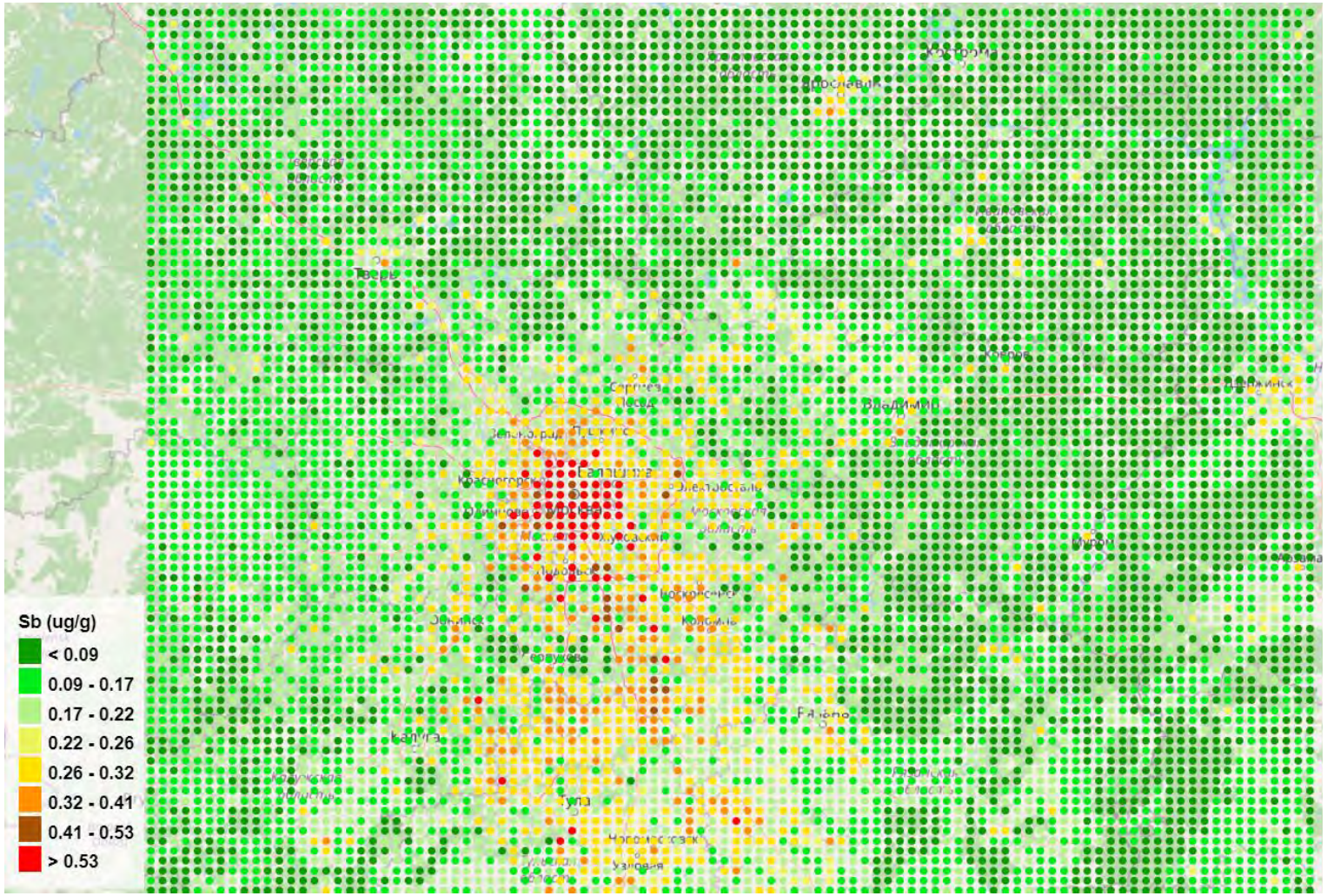
Results (2019 – 2020)



Sb contamination prediction for 2019 (left) and 2020 (right)

The lockdown in Russia that lasted for approximately 1.5 months imposed different limitations. Most of the limitations restricted the movement activities of the population. According to the official statistics, industrial production in Russia decreased by 2.9% from the past, by the end of 2020.

Results (High spatial resolution)



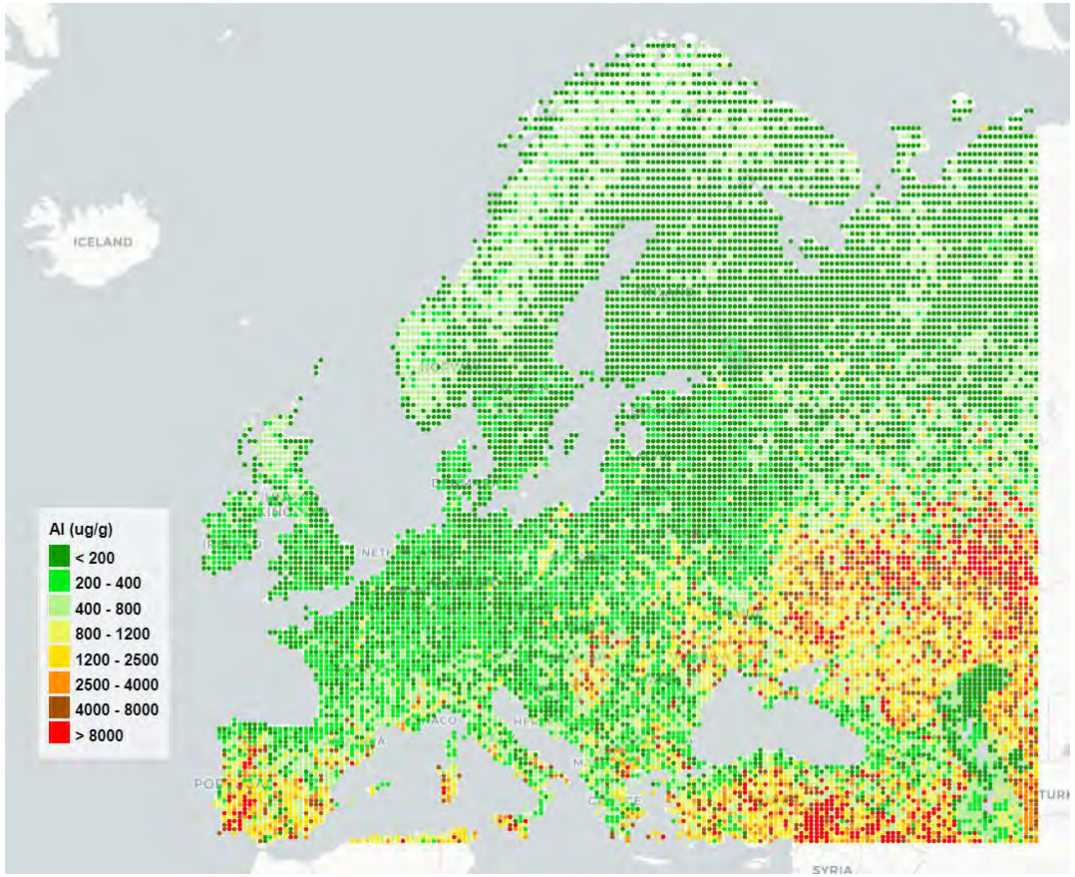
High spatial resolution of the SNN model prediction of Sb contamination

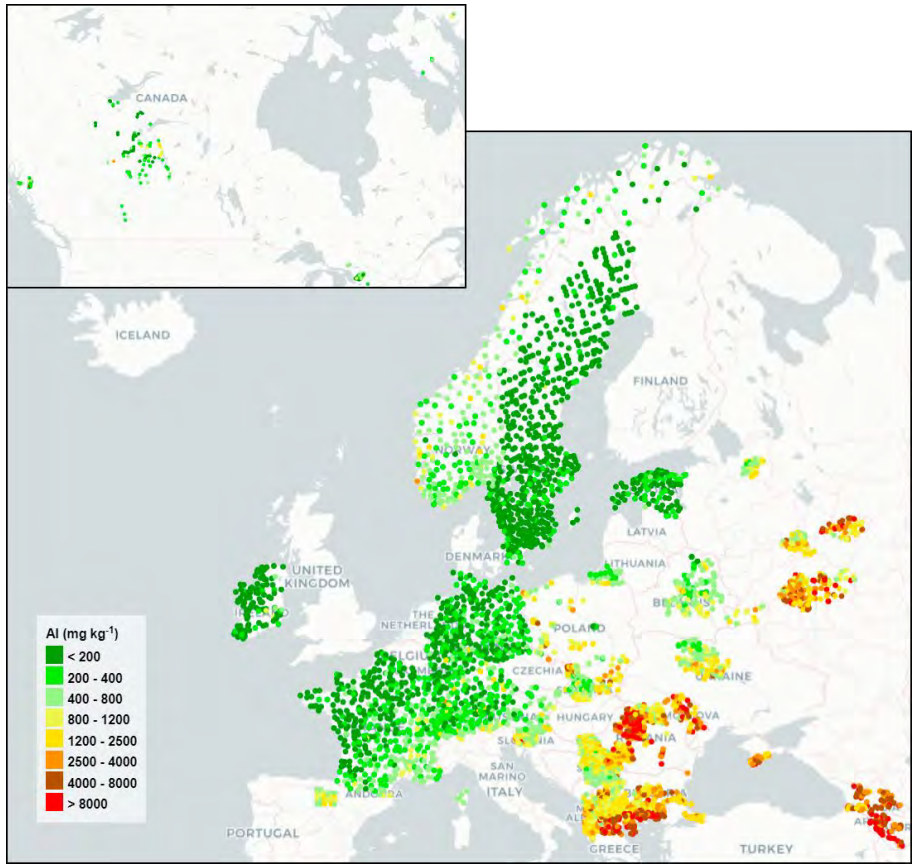
Moscow is a thickly populated city, and the population is increasing at a fast pace. Published information reveals, there are about 12.5 million habitants in Moscow. Therefore the Sb contamination level there is bound to be very high.

The map also reveals clusters of hot spots in large cities, such as Tula, Kaluga, Vladimir, Tver, Nizhny Novgorod, Yaroslavl, etc. It is also seen that from Sergiyev Posad to the north direction, the contamination level is rather low, except Yaroslavl, where the working oil refinery is located.

The Tula region stands out on the map. There is a multitude of industrial enterprises located in the region, i.e., chemical, metallurgical, and machine-building, besides several large thermal power plants. Huge transport nodes and federal freeways are seen, rather clearly, on the map.

Draft maps (model trained on Full data)





We are focused on regression and classification tasks; however, classification is prioritized since it becomes possible to apply balancing techniques for training datasets, and a gradation of pollution levels is initially used when building maps. For local and regional maps of some elements, the model accuracy reaches 90-95%.

Задачи:

1. Апробация различных статистических и нейросетевых моделей
2. Моделирование загрязнения PM2.5
3. Автоматизация процесса прогнозирования
4. Поиск альтернативных источники данных, чтобы проверить выдаваемые GEE значения.
5*. Расширение функциональных возможностей платформы с целью определения возможных источников загрязнения.

СПАСИБО, НА ЭТОМ ПОКА ВСЕ!