YANDEX SCHOOL OF DATA SCIENCE

Y-DATA Industry Projects

2019-2020

Yandex

Table of Contents

Industry Projects

Industry project at a glance:

Duration:

~5 months long

(February 2020 - June 2020)

Scope:

~350 work hours total

Team: 2-3 students

Ongoing guidance:

1 hr/week from Y-DATA

1 hr/week from company

Compatibility:

Projects are selected by students based on preference

Students approved by company for participation

Intro

The Industry Project is one of the key elements of Y-DATA program, translating theoretical knowledge into practical skills. We believe that in order to achieve full understanding of the use and application of machine learning algorithms, Y-DATA students should experience as part of their studies work on a real-life, full-cycle industry project. This experience will provide invaluable hands-on experience working on real data and problems and overcoming realistic challenges.

The process of working on the project follows popular industry standards and methodologies and incorporates a growing set of tools the students possess to methodically understand and solve a real-world problem. The work on the project is supervised and guided both by experienced mentors from Y-DATA, providing technical knowledge and understanding of the ML tools used, and by data owners from the company, offering domain expertise and understanding of the business aspects of the problem.

This catalogue presents the list of industry projects available for the 2019-2020 course year and background information on the companies offering them as well as the logistics for each project and company. The projects presented in this portfolio represent a broad range of domains and tasks, allowing each student to select projects suiting their interests and skills. Each project underwent a definition and refinement process in order to be suitable in scope and complexity for Y-DATA students' current (and upcoming) skillset and availability. The projects are all defined with a multi-tier goal structure, including a basic goal and one or more advanced goals. The basic goal for each project was chosen so we have high confidence in it being obtainable given determination and effort in the existing scope. The advanced goal(s) is more ambitious and more challenging, representing something to aim for, and which will yield significant benefit to the company if it is accomplished.

Project Selection:

After reading the full descriptions of projects available, each student should choose 5 projects from the list and mark them in order of priority in a dedicated form. We will strive to assign each student to a project they're interested in, taking into account the preferences submitted as well as the needs of a given project and company and the skills and availability of each student. In addition, some companies may be interested in having their representative talk to the prospective students before they're assigned to the project. The exact dates for the selection process appear in the timeline section of this catalogue.

Timeline

08.01.2020	Project catalog publication
15.01.2020	Deadline for submission of project priorities
25.01.2020	Publication of student assignment to projects
01.02.2020	Beginning of work on projects
02.2020	Data understanding, exploration, domain understanding
15.03.2020	Project presentation – Initial phase: Data, goals, tools
03-05.2020	Main work on project – modelling, evaluation etc
16-26.06.2020	Project results presentations (Internal)
TBD (Early July)	Demo day – public presentation of chosen projects

Academix

Academix

Domain:
Academic research discovery

Location:
Jerusalem

Intro:

Academix is a startup which helps companies and organizations seeking deep expertise and R&D capabilities in healthcare and the life sciences fields to find scientists from academia that meet their specific requirements. Academix is developing natural language processing (NLP) technology which maps out, quantifies and ranks expertise for the millions of academic scientists and medical researchers worldwide, enabling instantaneous identification of top experts in any field. Many of these developments rely on machine learning. The platform developed by Academix is being used by companies worldwide and government bodies in Israel to help expediate scientific discovery, and is recognized and supported by the Israeli Innovation Authority.

Logistics:

Work on the project can be conducted remotely, except for an initial meeting and summary meeting that will be held at the Academix's offices in Jerusalem.

- Academix will provide access to cloud resources required to conduct the task and access to all relevant data.
- Sign NDA, IP agreement.

#01

Academix

Type of task:
Clustering

Keywords:
Record Linkage,
Complex Network Analysis,
Graph Analysis,
Similarity

Disambiguation of researcher profiles using machine learning

Problem Description:

The scientific literature holds much of the human knowledge and innovation. Despite this, it is mostly unstructured, therefore it is currently impossible to accurately identify the body of work of a given researcher to credit their work and fully assess their expertise. For example, in Israel there are 3 active researchers named 'Nir Friedman', but how does one uncover which publications are authored by which Nir? This problem is part of a field of study named 'record linkage'.

Academix has developed proprietary technology to disambiguate between researchers based on weighing multiple inputs from scientific literature, taking into account that some features might be missing. While this technology produces accurate results in most cases, selecting the optimal similarity metric is challenging, and not always optimal. The aim of this project is to employ machine learning to conduct this clustering task and determine the optimal similarity metric in order to further improve the output and handle edge cases.

Dataset Description:

The dataset holds >30M publications that have been pre-processed by Academix's NLP algorithms and will be provided alongside two sets of cases which can be used for evaluating the project's success, the first set was generated internally and the other was published:.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930168/

https://figshare.com/articles/PLoS_2016_csv/3407461/1

Project Goals:

Basic goal:
Research the problem and create a model capable of disambiguating between researchers with comparable accuracy to the current algorithm. This includes finding the optimal similarity metric to measure distance between profiles.

Advanced goal:
Use complex network analysis and graph analysis tools to improve the basic result and to group individual researchers according to their distance in the network graph.

Anodot

Domain:
BA/BI

Location:
Raanana

Intro:

Anodot's mission is to turn Business Analytics from a manual process involving queries and visualization to autonomous analytics, producing insights and forecasts automatically. Anodot's core involves a large scale machine learning platform, developed by the company in the last four years. Anodot's first autonomous service is anomaly detection and root cause investigation, detecting and alerting on business issues at any level. Among Anodot's customers are Facebook, Waze, Tesla, TripAdvisor, Lyft, Wix, King, Microsoft, AT&T, Deutche Telekom, and many more. Anodot has 80 employees, in 4 offices (Raanana, Silicon Valley, London and Sydney), with R&D and Data Science all in the Israel office.

Machine Learning at Anodot:

Anodot currently has a team of 8 data scientists working on machine learning with time series data. Machine learning is the heart of the product and therefore any successful project will have a direct impact on the product and company.

Logistics

- On-site presence required for some of the time, most of the work can be done from home.
- Sign NDA, IP agreement (all IP produced will be owned by Anodot).
- Can be on own laptop, but should include data protection as the data is private and sensitive.

#02

Anodot

Type of task:
Research

Keywords:
Time Series Analysis,
Spatial Statistics

Massive time series - event correlation

Problem Description:

In the context of time series forecasting, in some cases, an event can have a huge impact on the behavior of a tracked metric. For example, around Black Friday or Christmas we observe a major increase in sales-related metrics. During a snowstorm, on the other hand, we can notice a significant drop in taxi use, etc.

Most of the time those events are known in advance and gaining an understanding of their impact and using it to integrate them into the model can dramatically improve the quality of the prediction. This understanding is known as event regressors - and it requires several instances of an event occuring (e.g, several snowstorm events are required to understand the impact of a snowstorm on the number of taxi rides taken in NY).

But given a huge quantity of events and a huge amount of time series, it is hard to find what is influencing what and which group of events have a similar impact. In particular, given a set of events that are tagged (e.g, sale events, software releases, sport events, etc) - the problem of discovering their impact on a time series has two parts:

a. Which groups of events have a similar impact on the time series being forecasted

b. What is their impact on the time series

While the second is solved with event regressors, it is tightly coupled to the first.

In addition, the typical scale of Anodot is around millions of time series and thousands of events, making it a scale challenge as well.

The goal of the project is to efficiently discover the group of events influencing time series.

Candidates should be able to conduct research from literature review, algorithm design, implementation and validation.

Dataset Description:

Several hundreds of thousands of time series along with several thousands of known events.

Project Goals:

Literature review of the problem

- **Basic:** An algorithm / method of how to find events that have an impact on each time series
- **Advanced:** Characterize the correlation between the events

Project Impact:

Improve the ability to accurately forecast time series metrics and improve anomaly detection at large scale.

AInovIA Labs

AInovIA

Domain:
Launchpad
for AI projects

Location:
HaSharon

Intro:

AinovIA Labs is an innovation startup launchpad established by an experienced entrepreneurial team. Through our work with partners and customers we identify market problems and gaps that can be addressed with AI. Our experienced AI teams develop relevant products & technologies that we commercialize or spin out as separate startups.

What’s unique about AInovIA?

- We work on real projects that we have customers for
- We work on exciting AI stuff at the forefront of AI research
- Highly experienced team that you can learn from
- Successful projects may spin-out as startups where you may play a leading role

ML at AInovIA :

AInovIA team has rich knowledge in the development of AI products in various fields, including DL, image/video and speech processing, edge AI, and more.

Logistics:

Working from home most of the time

Our team will provide continuous remote guidance with occasional F2F sessions when needed

Signing NDA and IP waiver is required

In most cases you can use your own laptop / computer

#03

AInovIA Labs

Type of task:
Speech Synthesis

Keywords:
Deep Learning,
Sound,
Scarce Samples,
Neural Networks

Neural Voice Cloning with Few Samples

Problem Description:

In speech synthesis, generative models can be conditioned on text and speaker identity. While text carries linguistic information and controls the content of the generated speech, speaker identity captures characteristics such as pitch, speech rate and accent. One approach for multi-speaker speech synthesis is to jointly train a generative model and speaker embeddings on triplets of text, audio and speaker identity.

The idea is to encode the speaker-dependent information with low-dimensional embeddings, while sharing the majority of the model parameters across all speakers. One limitation of such methods is that they can only generate speech for observed speakers during training. An intriguing task is to learn the voice of an unseen speaker from a few speech samples, a.k.a. voice cloning, which corresponds to few-shot generative modeling of speech conditioned on the speaker identity.

While a generative model can be trained from scratch with a large amount of audio samples, we focus on voice cloning of a new speaker with a few minutes or even few seconds data. It is challenging as the model has to learn the speaker characteristics from a very limited amount of data, and still generalize to unseen texts.

Dataset Description:

Data from tens of speakers in different accents with marked ground truth.

Project Goals:

- **Basic:**
  Improve existing convolutional network in several ways:
  Require less information
  Lower average loss functions on test dataset
- **Advanced:**
  Use a deep learning GAN approach

Project Impact:

Speech synthesis has long been a vital assistive technology tool and its application in this area is significant and widespread.

Improve perception - Speech-to-speech cloning will enable to improve our ability to understand people over the phone or internet calls, improve the ability of speech-to-text systems and improve the overall quality of experience.

Understanding speech cloning will enable to counter deep-fakes and develop new anti-fraud measures.

AInovIA Labs

Type of task:
Detection (in image)

Keywords:
Deep Learning,
Transform,
Face/Human/Figure Detection

Figure Detection in Near IR

Problem Description:

In recent decades a myriad of technologies including electronic toll collection and license plate recognition have been developed to improve the integrity of enhanced transportation systems. The focus of these technologies has mostly been on the vehicle instead of on the occupants.

An occupant detection system can have many practical applications, including: 1) facilitate the operation of high-occupancy vehicle (HOV) lanes, 2) collect data for transportation planning, and 3) monitor vehicles at public facilities, military bases or other sensitive sites.

In this project, we study a method to detect and count the occupants in a vehicle.

Dataset Description:

Multiple videos of moving cars captured in near IR with ground truth.

Project Goals:

- ■ Basic:
  - Build a face detection network for near IR images with over 90% accuracy
  - Build an image detection network for near IR images with over 90% accuracy
- ■ Advanced:
  - Achieve 95% accuracy for the NIR image/face detector
  - Achieve false positive rate of below 1%

Project Impact:

1. High Occupancy Vehicle/High Occupancy Tolling (HOV/HOT) lanes are operated based on voluntary HOV declarations by drivers. A majority of these declarations are wrong to leverage faster HOV lane speeds illegally. It is a herculean task to manually regulate HOV lanes and identify these violators. Therefore, an automated way of counting the number of people in a car is prudent for fair tolling and for violator detection.
2. While there is a lot of work in the market to handle face or figure detection in visible light, there is little to none work done on other spectrums. In recent years there is growing interest in NIR spectrum for a myriad of applications as automotive, safety, security, agriculture, industrial and more. The creation of CNN in NIR and the gained experience in this new domain will open new opportunities which far exceed the scope of this project.
3. Students working on this project will gain a unique experience in this new domain.

Artbrain

Artbrain

Domain:
Online Marketplace

Location:
Tel Aviv

Intro:

Artbrain's mission is to spread art around the world.

We partner with leading auction houses and art galleries and help to grow their business by building stronger customer relationships.

Our product is a SaaS solution for relationship marketing. We find the perfect items for each art collector and enable auction houses and art galleries to create and send personalized recommendations with the best items for each collector (via emails, phone calls, print catalogs, social media and online marketing campaigns).

We identify the most relevant items for each art collector and convert information overload into individualized professional knowledge, using human-based computation and machine learning, to separate and spread high-quality content.

ML in the company:

Artbrain uses third party NLP tools to analyze the Auction houses' art items. We recently started working on improving our NLP and recommendations system and we work with an experienced ML advisor that helps us to move quickly throughout the ML path. This advisor will guide you in this project.

Logistics:

Work location:
We can easily host the students at our offices, located at: Azrieli Sarona, 30th floor.

Legal requirements:
Sign NDA, IP agreement (all IP produced will be owned by Artbrain).

Data access:
Students will be expected to have their own laptops. Data will be provided as CSV.

#04

Artbrain

Type of task:

Recommendation\
Prediction

Keywords:

Recommender Systems,
User Engagement,
Clustering,
Transaction-Based Recommendation

Art Discovery Recommendation Engine

Problem Description:

We want to improve the quality of our recommendation.

Try to beat our current recommendation algorithm! Are you up to it?

- **Basic:** Beat our rule-based algorithm. Our first algorithm was developed to deliver a quick solution to our customers. It uses Google NLP service to identify the relevant keywords of each item's description and then matches these keywords with the user's history.
- **Advanced:** Beat our ML POC project. This is a new project that uses AWS ML tools to create the recommendations relevant for each user.

A fine whiskey bottle will be given for the team in case of success.

Dataset Description:

In our system, we have hundreds of thousands of past items and millions of interactions (submitted bids and purchases). The data is located in our DB and any variation of CSV files can be produced. The items are manually tagged in 2 levels of item department and category.

Evaluation:

We will perform A/B testing to evaluate the results, comparing the recommendation created by your algorithm with the existing recommendation. Our feedback is quick. If the recommended item was clicked on by the user, it is good. If the user placed a bid on this item, that's already a big thing.

Project impact:

In this project, you'll help art lovers to find their next art piece. Better recommendations will allow users to find exactly what they are interested in.

Data for Good

DATA FOR GOOD LAB

Domain:
Academic research

Location:
Be'er Sheva

Intro:

Our name - the Data Science for Social Good Lab - reflects our goal: to improve the world through data. A gigantic volume of data is now available in the world, and much can be accomplished if we attain and utilize it in an effective manner. Our aim is to make our research reproducible and open. We are constantly working on releasing various public large-scale datasets that include online social network datasets, such as Facebook, Google+, Academia.edu, and Reddit Communities; time series datasets; and the largest public network evolution dataset with over 20,000 networks and over a million real-world graphs. Our lab is a nonprofit academic centered lab, and we are a part of Ben-Gurion University data science center.

Logistics:

Work location:
Flexible, work wherever you want, though we can easily host the students at our lab, located at: Ben Gurion University building 96.

Legal requirements:
Publications/public statements that include material from the project will require the lab agreement (we are encouraging academic publications).

Data access:
Students will be expected to have their laptops. Data will be accessed either (1) directly on the lab servers, or (2) offline set will be provided.

#05

Type of task:
Image data analysis

Keywords:
Computer vision,
Complex networks,
Data exploration

Representation Of Real-World Events in Memes

Problem Description:

In the past century, the common ways to pass ideas and information changed drastically. It evolved from written media like newspapers to audio-based media like radio, then proceeded to visually-transmitted media such as television, and today millennials spread their perception and beliefs online in various new ways. One of these ways is memes, a sort of humorous image that spreads some ideas or concepts.

We believe that the study of the meme's life cycle can present many insights about millennials' opinions concerning real-world events. For instance, we can discover what millennials think about the gender gap problem, anti-vaccination movement, U.S elections, Tesla's Cybertruck, etc. We are planning to present a novel analysis that will map memes to real-world events and study how memes evolved alongside these events.

Dataset Description:

For starters, several thousand of image-based memes. In more advanced parts of the project we will provide larger datasets with additional metadata such as: comments, upvotes/downvotes, etc.

OH PLEASE
TELL ME MORE

MY DO TELL US WHY YOU HATE TRUMP AND ARE PROTESTING
YET YOU DONT SAY ANYTHING ABOUT THE 2 WORLD WARS YOU GRANDPARENTS MURDERED IN

IT'S VALENTINES DAY?
TELL US ABOUT HOW LONELY YOU ARE, WE ALL CARE SO MUCH.

Project Goals:

- **Basic:** Implement and analyze the results of a software package that receives an image meme data by performing the following tasks:
  - Text extraction using OCR.
  - Objects and their size extraction using/training segmentation, object detection models.
  - Performing automatic image captioning.
  - Returning similar memes by using a feature vector created by a CNN.
  - A social network of connected/similar memes.

An analysis should contain an exploration of the results of the package. For example, what objects are the most popular in memes, which is the most central meme by closeness centrality, TSNE of the embeddings of the image captions, etc.

- **Advanced:** Create a mapping between events in the real-world and fluctuations in meme lifecycle utilizing time series analysis tools. For instance, an interesting result could be to find the time shift between memes and events in the real world.

Project Impact:

This project will provide a novel technique for analyzing and getting insights from type of media that currently is very hard to analyze.

Dell EMC

DELLEMC

Domain:
IT infrastructure

Location:
Be’er Sheva

Intro:

Dell EMC is a part of the Dell Technologies family of brands. Dell Technologies is a unique family of businesses that provides the essential infrastructure for organizations to build their digital future, transform IT and protect their most important asset: information. The company services customers of all sizes across 180 countries – ranging from 98 percent of the Fortune 500 to individual consumers – with the industry’s most comprehensive and innovative portfolio from the edge to the core to the cloud.

Among DELL EMC specialties are information infrastructure, unified storage, content management, security, virtualization, backup and recovery, big data, Virtual Desktop Infrastructure, cloud computing, data federation, and deduplication.

Data Science in the Company:

The Data Science Services (DSS) team is located in Beer Sheva and currently includes 10 team members from diversified scientific backgrounds. The team charter spans the entire Data Science lifecycle, including: define the business problem, access and clean the data, build and train AI models leveraging domain expertise and support implementation in the customer environment. Most projects are performed for internal business units, for example finance, supply chain and engineering units developing DELL EMC core products. In the last couple of years, the team was also involved with external DS consulting projects for some of DELL EMC strategic customers.

Logistics:

- On-site presence required for some of the time, most of the work can be done from home.
- Sign NDA, IP agreement (all IP produced will be owned by DELL EMC).
- Can be on own laptop, but should include data protection.

#06

Dell EMC

Type of task:
Anomaly Detection

Keywords:
Sensors Data,
Multivariate Time Series,
Anomalies,
Outlier Detection

Dynamic Anomaly Detection in Complex IT Environments

Problem Description:

The Data Protection Division develops tools to help clients protect their data. One key feature/technology is the ability to protect virtual machines data by saving it on a cloud.

Moreover, in case of a disaster, it is possible to restore the VM on a cloud machine. The snapshots of VMs can be stored both locally and on a cloud. In case of a disaster a cloud snapshot can be used to create a VM on Aws/Azure/VC. The system contains two main services, one is responsible for sending snapshots to the cloud, and a second service that runs on the cloud and is mainly responsible for orchestrating the recovery of snapshot back to a running VM. The described system is complex, and it is hard to monitor and detect problems immediately. Solving a problem once it is found, also requires a tedious process of searching for what caused the problem. Developing a data-driven approach to accomplish both will be highly beneficial.

Dataset Description:

Datasets comprised of aggregated minute/hourly/daily data of numerical KPIs and also logs data. Numerical KPIs can be: Snapshot size change rate, Snapshot upload rate, Number of protected VMs, Number of protection policies which does not meet SLA, CPU, Memory, Number of Threads and more.

Project Goals:

- **Basic goal:**
  Semi-Supervised anomaly detection - detect historical anomalies in a complex environment, which will be evaluated by domain expert.
- **Advanced:**
  Perform causal inference on historical anomalies to detect presumed cause, and use insights to create a system for live anomaly detection.

Project Impact:

The impact for this project is the ability to recognize a problem in the system so that a customer can proactively take action as fast as possible in order to prevent it from happening.

Dell EMC

Type of task:
Classification

Keywords:
Sensor Data,
Multivariate Time Series,
Failure Prediction

HDD Failure Prediction

Problem Description:

HDDs and other storage devices produce data that can be collected and used for monitoring and alerting. Dell collects such data for various uses, one of them is HDD failure prediction. The ability to alert beforehand that a drive is going to fail can be valuable for customers and hence for Dell. In this project we use daily data collected over a long period of time from HDDs in Dell's machines. This data can be treated as a multivariate time series. The goal of the project is to build models using ML&DL that will be successful in predicting that a drive is going to fail a few weeks before it does.

Dataset Description:

Dataset comprised of daily data of HDD attributes like S.M.A.R.T, temperature, etc. for several tens of thousands of drives. The labels are an indication of failure, and appear in about 1:100 ratio.

Project Goals:

- **Basic goal:**
  Execute simple feature engineering and implement at least one method of classification that will predict which drive is going to fail in the next 14 days.
- **Advanced Goal:**
  Execute sophisticated feature engineering and test several classical ML models, and/or develop DL models to improve the accuracy of prediction.

Project Impact:

The impact for this project is the ability to recognize a risk of failure to a drive, so Dell can send a field engineering to replace the drive on time. This will reduce the chance customers will suffer from data unavailability. This can also reduce costs by planning better the field engineer work.

DoubleVerify

Domain:
AdTech

Location:
Tel Aviv

Intro:

DoubleVerify is a data, analytics and measurements company for online ads - think of us as Google Analytics for ads. We track and analyze tens of billions of ads every day for the biggest brands in the world like AT&T, Vodafone, Disney and most of the Fortune 500 companies, across all platforms; Desktop, Mobile, YouTube, Facebook, Connected TVs, etc.

We do everything in real time to provide both analytics and the ability to preemptively block the ad from appearing in scenarios the advertiser wants to avoid (such as an ad for an airline company running next to a story about an aviation disaster, or an ad served to a bot mimicking a user behavior).

We provide a wide array of analytical products, here are examples for three major ones:

Fraud Protection - analyzing and detecting a wide range of online fraud schemes like automated bot behavior that mimics user activity and is used to browse sites and generate fake revenue, malware that injects ads to users' browsing, apps that run ads in the background without anyone knowing, emulators and VMs impersonating various devices, sites that buy popup traffic, and more.

Brand Safety - analyzing the content next to which ads appear and help advertisers avoid certain content like that isn't aligned with their brand like; Disasters, Fake News, Racism, etc.

Engagement - track users' engagement with the ads including viewability metrics, clicks, mutes, skips, etc. and help brands to optimize their campaign performance by understanding what attracts users about their ads.

Data Science in DoubleVerify:

We are a data and analytics company, and as such data is our main focus. We currently utilize Machine Learning mainly for fraud detection, but plan to expand that usage to many other fields and products, as can be seen in this project

Logistics:

- Working on this project does not require being in DV's offices.
- We'll be able to provide consultation, answer questions and give advice on a continuous basis either by email or by coordinating meetings.
- Signing NDA and IP waiver is required.
- Work can be done on local laptop, but should maintain data protection due to the private nature of the data.

#07

DoubleVerify

Type of task:
Classification,
Scoring,
Clustering

Keywords:
Ad Fraud,
Mobile Apps,
Classification,
Scoring,
Clustering,
NLP

Detecting Ad Fraud in Mobile Apps

Problem Description:

Ad Fraud is a major problem plaguing the industry, with mobile apps being especially susceptible to ad fraud due to the devices being in constant connection to the internet as well as the ease with which anyone can get an app on the Play Store. Ad fraud includes such behaviors as apps that run ads in the background, draining your battery and data plan; apps that generate pop up ads, or display them on top of other apps, and additional similar behaviors - all very frustrating for the user.

Given that there are millions of apps in Google Play Store, tens of thousands of monthly updates, and thousands of new apps are submitted daily to Google Play Store, it becomes a challenging task to detect which of these apps are exhibiting fraudulent activity.

We would like to score (and eventually classify) apps based on their metadata from the Play Store, and possibly additional data sources, with emphasis on user reviews.

Dataset Description:

The dataset provided will include all available metadata for apps from the Play Store, including user scores, user reviews, sentiment analysis labels for written reviews, requested permissions, category, number of downloads, developer details, and more. This will be supplemented by information on which apps were classified as fraudulent via manual review., containing several thousand unique entries. Additional lists of fraudulent apps can be obtained from other sources with our help and guidance.

Projects Goals:

Basic:

- Design and build a classifier that will detect apps with fraudulent behavior based on the metadata collected from the Play Store. The initial model features will include the already provided sentiment labels for the user reviews. Additional features should be defined during the work.

Advanced:

- Improve the basic model further using topic modeling, seeking out relevant reviews that describe fraudulent activity (ads in the background, pop up ads, etc.)
- Enrich and expand the model with additional collected labeled data and improve the classifier based on that data.

Project impact

The model will be evaluated and if accuracy is satisfactory it will help us to prioritize which apps are to be analyzed further.

Dynamic Yield

Domain:
Marketing Tech/
personalization

Location:
Tel Aviv

Intro:

Dynamic Yield is a personalization company that powers individualized experiences for more than 600 million users each month across hundreds of global brands.

Its product gives Marketers the tools to fully personalize their website: Recommendation Widgets with various strategies, Experimentation and A\B testing, Personalized Email Campaigns and Push Notifications, User Segmentation based on custom or machine created rules and more.

ML in DY:

We use ML for various tasks. Predictive Segments (populations that are likely to convert), Product Recommendations, Automatic Variation selection (Using Multi-Armed Bandits) and more.

As a multi-tenant SAAS company, one of the challenges we have to tackle is training and predicting many per-customer models simultaneously., in low latency.

For that purpose, we developed an infrastructure that abstracts the different stages of the model lifecycle, from preparing data to serving the model in real-time.

We have a team of 3 Data Scientists, and a few engineers specializing in ML.

Logistics:

- We prefer students to work on-site at our cool Tel-Aviv office at Igal Alon for at least once a week, and from home. Some flexibility will be possible based on personal needs.
- Sign NDA, IP agreement
- Work on company provided laptops

#08

Dynamic Yield

Type of task:
Recommender Systems

Keywords:
NLP,
Word2Vec,
Nearest Neighbours

Deep Learning Product Recommendations

Problem Description:

One of the key features in DY's offering is product recommendation widgets.

These recommendation widgets come in various strategies, both personalized (based on user's history) and not (top products, bought together).

The latest strategy that was implemented is a naive NLP-based Deep Learning strategy, which treats browsing history as a "story" of the products the user interacted with.

An NLP (word2vec) neural net is then trained on these "stories", yielding a model that can recommend which products to show to a user to maximize purchase probability in real-time.

The starting point of our project will be this preliminary algorithm, which is still in incubation.

Topics for research and implementation will be: different optimization goals, robustness of evaluation, minimizing bias in training data, adding more data points (sessions, product attributes, recency decay etc.), optimization of the predict step and more.

Dataset Description:

Anonymized behavioural user data from major E-Commerce sites (all user interactions with the site: product views, clicks, e-commerce events as purchases and add to carts),

In addition to the site's product catalog with product attributes such as price, brand etc.

Project Goals:

- **Basic:**
  Research and evaluation of different model\inference improvements against baseline. Show an improvement over the current model.
- **Advanced:**
  Evaluate and implement a joint model that uses more than one technique to further improve against the baseline.

Project Impact:

Millions of customers in leading e-commerce sites will be exposed to the improved product recommendations.

Evogene

Domain:
Plant genomics

Location:
Rehovot

Intro:

Evogene is a leading biotechnology company developing novel products for life science markets through the use of a unique computational predictive biology ("CPB") platform.

This computational platform relies on deep multi-disciplinary know-how in biology and chemistry, combined with cutting edge computational technologies that allows the creation, integration and analysis of dedicated Big-Data for the purpose of systematic discoveries in the chosen areas of focus. This unique technology has evolved over the last decade and has been enhanced during the years with technological advancements in Artificial Intelligence, Deep Learning and Big Data.

Today, the platform is implemented in three general life-science based markets: agriculture, human health and life-science based industrial applications.

Logistics:

Work location
We prefer students to work on site most of the time but partial work from home is also possible. Minimal time of a few hours per week on site is crucial for project success.

Legal requirements
An agreement will be signed between Evogene, Yandex and the students

Data access format
We prefer students come with their own laptops to potentially serve as terminals for our computational system. Precise details to be determined.

Background needed
Students should have prior knowledge in biology, at least in B.Sc level

#09

Evogene

Type of task:
Binary Classification

Keywords:
Unbalanced Dataset,
Supervised Learning,
Data Driven Biology

Prediction of bacterial gene functions

Problem Description:

Bacterial genes are one of the most diverse sources for enzymatic activity or other types of molecular functions in the world. Over the last few years Evogene has successfully discovered specific bacteria and bacterial genes for various agricultural and human health applications. However, in many cases we identify genes with unknown functions, limiting our ability to further develop the product based on these genes. Over the past several years we have designed unique computational descriptors for bacterial genes, compiling various aspects of the gene itself and its context within the genome. We would like to use these gene features along with a training set of genes with a known function to be able to predict the function of yet uncharacterized bacterial genes.

Dataset Description:

- The training set is comprised of tens of thousands of genes classified based on the presence of a function of interest into positives and negatives
- Number of features – up to 30,000, separated into groups.
- The training set is highly unbalanced - the number of negatives can be orders of magnitude higher

Project Goals:

- **Basic goal:** Develop a classifier that will predict whether a new gene possesses the function of interest or not.
- **Advanced goal:** Expand the system to incorporate feature importance understanding in order to reliably reflect the importance of the different features and sets of features for a successful classification of future genes.

Project Impact for the company/product:

If successful, the project's results will significantly boost the ability to use yet-untapped genes for practical applications in the life science domain including agricultural, human health and industrial applications.

Explorium

Domain:
BA/BI

Location:
Tel Aviv

EXPLORIUM

Intro:

Explorium is an automated data and feature discovery platform.

Explorium helps data scientists and organizations improve the performance, explainability, and accuracy of their predictive models in a fraction of the time by automatically discovering and applying more relevant data and high impact features for best-in-class ML algorithms.

Our product has five major components: data discovery, feature extraction, feature selection, model training, and model serving.

Data science in the company:

We are a data science company, so in one way or another we are all focused on data science and our workflow is all about machine learning.

Approximately half of the R&D team at Explorium is 100% data science-focused, with various levels of expertise, among which are:

- Research Team
- Internal Data Science Team
- External Data Science Team (customer-facing)
- Data & ML Engineering Team (data pipelines, creative engine)

Logistics:

- Location - on-site or at home (both are ok)
- Legal - students will sign NDA upon starting, IP agreement
- Data Access - Own student’s computers

#10

Explorium

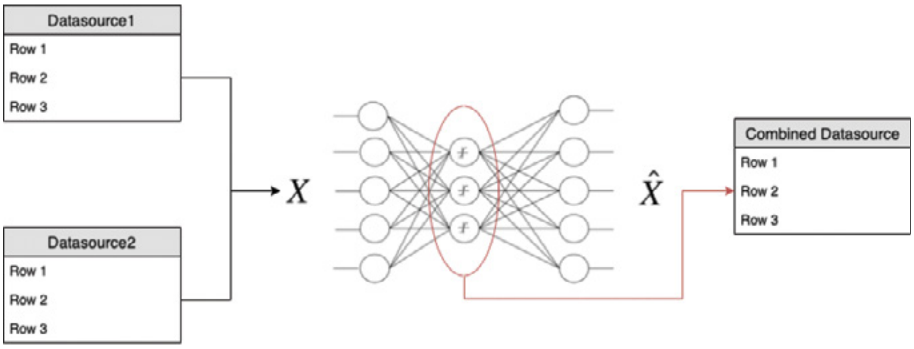
Type of task:
Feature Generation

Keywords:
Self Supervised Learning,
Autoencoders (AE),
DNN,
RNN,
CNN,
PCA,
SVD,
ICA,
Embedding,
Imputing

Self-Supervised Feature Generation

Problem Description:

Explorium has an extensive amount of data sources. We aim to create new data signals by combining multiple data sources and using various techniques from linear models to deep learning, we will then use the new signals to train models and test them against model trained on the original data. Such new signals will allow us to provide reliable data without compromising the privacy of sensitive data while retaining the quality of the models.



Dataset Description:

Unlabeled high-quality data with ~1M rows with ~100 columns and 2 indexes over 2 different tables and an API + multiple labeled datasets of several business tasks for the evaluation process.

Project Goals:

- **Basic:** Working linear feature generator that (over-)achieves original-data model accuracy. Working with numeric data
- **Advanced:** Non-linear feature generator and non-numeric data
- **Stretch Goal:** Using advanced methods to create/impute new samples (TL / GANs).
- **Evaluation:** we will evaluate the model in multiple ways.
  a. Signals vs. original data over the dataset used in creating the autoencoder (AE) in multiple classification tasks (real customer data).
  b. Signals vs. original data over an unseen dataset in the same ML tasks.
  c. Linear signals vs. AE signals over in the same ML tasks.
- Metrics that will be used are mainly AUC, R-squared, precision and recall.
- An AE model is the main deliverable for this project - it is of high priority to be able to retrain the model with the existence of new data, new activation function, or new initial weights.

Project Impact:

Creating private signals have impact on Explorium's ability to provide reliable data without compromising the privacy of sensitive data and the quality of our customers' machine learning models. If successful, the project will be embedded in the company's core data pipeline and will be used to create new enrichments that will immediately impact real machine learning models in production systems

Fermata

FERMATA

Domain:
Ag-Tech

Location:
Tel Aviv

Intro:

Fermata is a start-up that develops data-driven solutions for agriculture. Our products include tools for early disease identification, yield prediction, and cost optimization both for fields and greenhouses.

Our expertise covers data collection and data analysis. We develop a system for real-time plant monitoring using visual data to track plant growth and health, and combine this data with climate parameters using satellites, sensors and robots developed by our team. All the data is processed to develop recommendations for the farmers according to their needs.

Though Fermata was established only in early 2019, we already have numerous clients in Israel and Europe, including food retailers, technology developers and agricultural manufacturers. Fermata won AgriTech startup competition and is supported by Nvidia Inception program.

ML in Fermata:

Machine learning is the main focus of Fermata. We use Deep Learning for image analysis to monitor plants and identify diseases. We also use classical Machine Learning for yield prediction and to develop the models for cost optimization.

Logistics:

Work location:
Work on the project can be conducted remotely

Legal requirements:
Sign NDA, IP agreement

Data access:
Students will be expected to have their own laptops

#11

Fermata

Type of task:
Image Classification

Keywords:
Computer Vision,
Image Classification,
Agriculture,
Multiclass Classification

Plant species and disease state classification

Problem Description:

One of the key tasks existing in agriculture is analysis of plants' state and early detection of diseases. However, this task requires a large number of people with special technical skills. For example, a standard tomato greenhouse requires about 300 staff members for plant-state monitoring. The goal of this project is to build a system capable of differentiating between healthy and diseased plants species/varieties and, potentially, identify specific diseases.

Dataset Description:

We will provide several separate datasets of different plants/varieties tagged with known health status (healthy/diseased):

(1) ~9k proprietary images of smaller greenery, with a general label of healthy/diseased. (2) Large (~30k) database of low-resolution images of plant leaves from various species labelled as healthy or with a known disease. (3) Small (~1k) dataset of single species labeled by experts with as healthy or with 2 specific diseases. (see examples below)

Project Goals:

The main goal of the project is to develop an algorithm for determining the health status of a plant based on an image. The specific goals are as follows:

- **Basic:** build an algorithm for binary classification between 'healthy' and 'unhealthy' classes in "Satin" Lettuce from Dataset #1 and achieve maximum recall (≥ 0.7) while maintaining precision > 0.5. If needed, all the images from all datasets can be used in this task.
- **Advanced 1:** compare different strategies for classification and recommend best practice for future tasks: should we use two classifiers, with the first used to detect species and the second making health prediction for the specific species or one general classifier with paired outcomes 'species + healthstate'.
- **Advanced 2:** build a multiclass classifier for dataset #3 capable of differentiating between healthy state and two diseased states.

Project Impact:

Currently agricultural manufacturers lose up to 30% of their yield due to plant diseases. This project will help Fermata to build the solution for early plant disease diagnostic.

Fiverr

fiverr

Domain:
Online marketplace

Location:
Tel Aviv

Intro:

Fiverr is an online marketplace for freelance services. Our mission is to shape the future of work. Founded in 2010, the company HQ is based in Tel Aviv and provides a platform for freelancers to offer services to customers worldwide. Fiverr provides services in more than 300 different subcategories. Fiverr has 400 employees in 6 offices

(Tel Aviv, New York, San Francisco, Florida, London, and Berlin). R&D & Data Science teams are located in Tel Aviv.

Data @ Fiverr:

Fiverr has four teams of product/business analysts and two teams of data scientists. Our data science team works on a variety of problems such as recommender systems, spam detection, time series, NLP, search and more. Fiverr is a data-driven company and data science teams are involved in every step of the way.

Logistics:

- Sign NDA, IP agreement (all IP produced will be owned by Fiverr)
- Work on site (TLV office) or from home
- Work can be done on your own laptop

#12

Fiverr

Type of task:
Regression

Keywords:
Linear Regression,
Logistic Regression,
Decision Tree,
NLP,
BOW

Gigs LTV Prediction

Problem Description:

There are many Gigs offered on the platform (a Gig is a service offered on Fiverr). Prioritizing different Gigs in the catalog can lead to a higher overall revenue and can also impact the success or failure of a specific seller. When new Gigs are created on the platform, they are even harder to prioritize, since we have no data on the new Gig.

Predicting the impact of each Gig can help Fiverr increase revenue and help new sellers to succeed.

What is the impact of a Gig? What is considered good/bad seller? We are going to figure out all of these together :)

Dataset Description:

There are a few different types of data we can work with:

1. Gigs - title, description, pic, image, price, lead time and more...
2. Seller – Geo location, seller description, education, skills and more...
3. Listing - CTR, impressions, Gig positions and more...
4. Orders - purchases history of the Gigs

Project Goals:

- **Basic:** Predicting the LTV of existing Gigs (Gigs with history data on the platform)
- **Advanced:** Predicting the LTV on new Gigs in the platform
- **Evaluation:** RMSE/ MAPE will be used as evaluation metric

Project Impact:

If successful, the main impact of this project is increasing Fiverr’s revenue

We also expect to see the following:

- Decrease the time users spend searching for the right Gig
- Increase the variety of new sellers on the platform

Gong

Domain:
Market intelligence

Location:
Ramat Gan

Intro:

Gong.io helps salespeople and other customer-facing roles have better conversations by using AI and machine learning to automate big parts of their work and coach them using real data.

Gong is a well-funded, high-growth start-up, which proudly serves hundreds of customers, including Facebook, Salesforce, PayPal, LinkedIn, ZipRecruiter, and more.

We have access to huge amounts of data including recorded voice conversations, written documents and CRM data. We use this rich data to build all of our models in-house, from ASR (Automatic Speech Recognition) to Video and Audio analysis, NLP and statistical modeling.

Machine Learning at Gong:

The Gong research team includes 10 very experienced researchers, working on a large array of speech, computer vision, NLP and statistical inference problems. Gong is recognized as a leader in AI and ML in Israel, and in the fields of NLP and Speech technologies in general. Machine learning is at the heart of the Gong platform and the proposed projects all have clear impact on the Gong product.

Logistics:

1. On-site presence required for some of the time, most of the work can be done from home.
2. Sign NDA, IP agreement (all IP produced will be owned by Gong).
3. Gong will provide a laptop and cloud machines for the project.

#13

Gong

Type of task:
Machine Translation

Keywords:
Audio Data,
Deep Learning,
Speech Recognition,
CTC,
Speech Transcription

End-to-end Deep Learning Speech Recognition Platform

Problem Description:

Automatic Speech Recognition (ASR, aka Speech-to-Text), is historically comprised of two parts – an Acoustic Model, mapping speech to phonemes, and a Language Model, mapping phonemes to words. While in the fields of Computer Vision and NLP, deep learning models outperform other approaches by large margins, in the field of ASR the vast majority of commercial systems don't use deep learning.

In recent years, there have been reports on success in applying end-to-end (e2e) deep learning architectures to transcribe speech. E2E DL platforms hold the promise of better adaptation to accents or new languages, making the use of phonetic lexicons obsolete. Two promising directions are Nvidia's Jasper and Facebook's Wav2Letter architectures.

Dataset Description:

Gong will provide a training set of 500 hours of English sales calls, manually transcribed, and manually transcribed test sets of 10 hours in each language. A publicly available dataset of Spanish will also be provided.

Project Goals:

- **Basic goal:** Develop a Deep Learning based system for Automatic Speech Recognition in Spanish. Given an audio wav file, the system should provide its transcription in a factor of approx. x5 real time, along with time stamps for each transcribed word.
- **Advanced goal:** Given a test set of 10+ hours of sales calls from the Gong data in each language, reach Word Error Rate (WER) that is lower than that of the Amazon Comprehend ASR system in the same language.

Project Impact:

The project will allow Gong to improve its service in Spanish and English. The quality of transcription affects the quality of downstream tasks as well, including NLP understanding of the conversations. Gong's existing Speech Recognition system does not use Deep Learning, and reaches a very low WER compared with commercial solution like Google or Amazon, and is key to Gong's success in Natural Language Understanding. Having a Deep Learning based network for ASR in some languages will allow Gong to more easily support additional languages by training them on relevant data.

Gong

Type of task:
OCR

Keywords:
Image Segmentation,
Detection,
Object Localization,
Semantic Similarity,
Image Embedding

Optical character recognition for slides and URLs

Problem Description:

During sales calls, the Gong platform captures screen shares and detects whether the screen share is of a browser demo or a presentation (e.g. PowerPoint). When a presentation is given, we want to retrieve its title (text), to allow easy navigation between slides and to find similar slides in other calls. This isn't simple, as presentation slides often include logos or large graphics, weird fonts and complex layouts, where off-the-shelf OCR platforms often fail. The task would use the open-source Google Tesseract tool that employs LSTM networks for OCR, and can be trained for the custom domain.

In browser mode, it is also beneficial to detect the URL of the page that is shown for simpler navigation and labeling. The problem is that the URL text is very small and might require image enhancement techniques for proper retrieval and that users use different browsers (Chrome, Firefox, Safari, etc) in different locations of the screen. The task here is to build a custom object detector using CNN for fast detection and localization of the browser bar and to enhance the OCR algorithm for proper detection of the browser URL.

For all images, it is often useful to find similar images in other calls. Among other things, this can be used to link to similar discussions by top salespeople for better coaching, or to aggregate top questions asked by customers during the display of a certain slide (e.g. "Pricing Plans"). If OCR performs well, it can be done using the extracted text. Alternatively, methods of image similarity using DL embedding can be used.

Dataset Description:

Gong will provide a labeled dataset of images along with their textual content, and ranks of similar/dissimilar items for project evaluation. Gong will also provide labels of bounding boxes for browser URL bars and the URLs in those bounding boxes.

Project Goals:

- **Basic Goal:**
  a. Extract the title from images of presentation slides using OCR
  b. Detect the browser URL tab in images and extract the current URL
- **Advanced Goal:** Train an algorithm for retrieving similar images based on semantic similarity using DL embedding
- **Stretch Goal:** Implement an effective image retrieval platform that can efficiently index millions of images, so that given an image the platform will return a ranked list of similar images in a matter of seconds

Project Impact:

The project will allow Gong to provide navigation cues showing where different slides are presented in a call. For example, a user would be able to quickly find calls where the "Pricing Plan" slide was shown for a long time, jump to the exact part of the call in which the slide was presented, and see links to other calls by top salespeople to see how they present the topic.

Healthy.io

Healthy.io

Domain:
Healthcare

Location:
Tel Aviv

Intro:

Healthy.io is one of the first companies to successfully turn a smartphone into a clinical-grade medical device. It has pioneered the category of smartphone Urinalysis, offering an FDA-cleared home urine test equivalent to lab-based devices. To achieve lab-like quality on commodity cameras, our multidisciplinary team deals with complex challenges in the intersection of machine learning, computer vision, clinical practices, and product design. Today, Healthy's solution has already helped thousands of women access faster treatment for UTIs and aided in the diagnosis of hundreds of previously undetected chronic kidney disease cases. Our solutions go beyond urine analysis. By bringing Healthy's unique expertise in AI and computer vision to the wound-care domain, we are able to provide innovative technology for digitizing, measuring, and tracking the progress of wounds, revolutionizing the way chronic wounds are managed and treated.

Deep Learning at Healthy.io:

Healthy has a team of algorithm developers and researchers working on a mix of deep learning and computer vision algorithms to analyze large-scale imagery and video datasets in the pursuit of making medical diagnostic widely accessible. Machine learning is at the core of Healthy's products, from object detection and color classification of dipstick images, to accurate segmentation and tracking of skin wounds in 2D images.

Logistics:

- Work on-site at our beautiful Tel-Aviv office (presence required), and from home.
- Sign NDA and IP waiver.
- Can use local laptop, but should maintain data protection due to the private nature of the data.

#14

Healthy.io

Type of task:
Classification

Keywords:
Machine Learning For Computer Vision,
Few-Shot,
Anomaly Detection

Classification of dip-stick images

Problem Description:

Urine examination using a dipstick has long been used to determine the health status of a person. A standard urine test strip, or dipstick, is comprised of multiple (different) chemical pads or reagents. These pads react and change color when immersed in a urine sample. Healthy's app captures (scans) images of a dipstick together with a reference color board and compares them using machine vision, to derive clinical values of the different parameters. Analyzing images taken by lay users in a variety of environments poses a significant challenge - image quality and stick placement and orientation may vary significantly, impacting the robustness of color detection algorithms and results accuracy.

In this project, your goal is to build a deep learning model that detects bad scans, images with problematic stick situations such as missing stick, partial visibility, bad orientation, wrong stick (and more), and classifies the type of problem observed. The model can be used to improve user experience, e.g., prompt for a retry, and the performance of our system

Dataset Description:

You will be provided with a crowd-sourced dataset of images of urine stick scans taken by users, with labels describing whether the scan is valid or invalid, and the type of problem, overall five types of errors (missing stick, bad orientation, etc.)

Project Goals:

- **Basic goal:** The basic task is to classify the type of scan (an image) as either valid or invalid, as well as learning the type of the error. The classification quality will be evaluated using a confusion matrix and a combined score based on both precision and recall. As part of this goal, you would also be tasked with capturing additional images to improve the diversity and balance of the data, and consequently the quality of the derived model.
- **Advanced goal:** Anomaly detection. Even with a large scan dataset, a derived model remains limited in certain cases - changes in user behavior or the environment can produce images with unexpected features. Our goal is to build a model that detects new types of problematic scans, that were not (or hardly) observed before. Solutions approaches to consider are few-shot learning (the ability to learn from few labeled samples), data augmentation, and classical anomaly detection techniques, among others.

Project Impact:

Urinalysis is one of the most common methods of medical diagnosis. Detecting problems in urine scans is a fundamental task for building a robust analysis algorithm. This project has a significant impact: it allows us to better understand user behavior and scan quality in the wild and support a much wider range of settings, making early detection accessible to (almost) everyone.

Healthy.io

Type of task:

Image Segmentation,

Video Object Segmentation,

Semi-Supervised Learning

Segmentation of body parts in images and videos

Problem Description:

Chronic wounds are a critical issue believed to affect as much as two per cent of the population. Accurate wound measurement and monitoring are crucial for effective wound care, but traditional manually-based measuring techniques are imprecise and can cause discomfort. Healthy's solution leverages a smartphone and machine vision algorithms to digitize wounds from multiple viewpoints and accurately measure and track their progress, to improve the way wounds are treated and managed.

A key task of our system is to segment the wound from an image. It is the basis of wound tissue analysis, and it is used to measure the area of the wound directly. A significant challenge for effective segmentation is the need to remove environmental background (such as surfaces, clothes, etc). Background removal highlights the wound features, augments the data, and simplifies the segmentation task. In this project, your goal is to develop a deep learning model for skin segmentation, often implemented using CNNs, that seperates foreground (skin and wound) and environmental background in 2D images.

Dataset Description:

A large-scale imagery dataset of body parts with ground truth segmentation masks.

Project Goals:

- **Basic goal:** develop a deep learning network for skin segmentation using the dataset, separating skin (inc. wound) and background. Model evaluation: using standard metrics for segmentation such as intersection-over-union and dice coefficient.
- **Advanced goal:** Video segmentation. Develop a deep learning network that learns to segment a video of a body part using only a small subset of labeled frames. Such a network can significantly augment the training data, i.e., in a semi-supervised fashion; and can be viewed as an extension of image segmentation across temporal units to propagate spatial info over time.

Project Impact:

Accurate detection of foreground and background regions in 2D medical images simplifies the task of object segmentation, highlights wound features, and enables accurate measurements, thus improving digital wound care management.

ICL

Domain:
Manufacturing

Location:
South Region

Intro:

ICL, a global manufacturer of products based on unique minerals, fulfills humanity's essential needs, primarily in three markets: agriculture, food and engineered materials.

The experience, knowledge and professionalism derived from establishing Israel's potash industry over a period of 80 years of intensive activity and major investment in R&D, have transformed ICL into a world leader in specialty fertilizers, bromine and flame retardants. ICL produces approximately a third of the world's bromine, and is the world's sixth largest potash producer, as well as one of its leading providers of pure phosphoric acid.

ML in ICL:

As part of the Industry 4.0 transition, ICL implementing IOT & Machine learning for predicting equipment failures and for improving the way we operate our production processes. Machine learning, with the guidance of Ben Gurion University – Software & Information system engineering department, is becoming to be important tool in ICL engineer's toolbox.

Logistics:

In order to secure ICL production information, each student that take part in the ICL project will need to sign NDA form. Project workplace will be in Beer Sheva (University or ICL HQ). The project team will be guided by 2 of ICL engineers (for any information & question answering).

#15

ICL

Type of task:
Anomaly Detection,
Prediction

Keywords:
Anomaly Detection,
Time Series,
Prediction,
Predictive Maintenance,
Power Failure,
Sensor Data

Predicting electrical power failures

Problem Description:

ICL has a major potash, bromine & magnesium production side located in Sdom Israel. This production site has two electric substations – north and south. The north substation is supplying electricity for our barges harvesting carnallite & feeding the site plants with raw material. In the last 3 years we are having several power supply failures resulting in production losses. We would like to predict those failures in order to offer a solution in such a way the that power distribution remains continuous and production losses are avoided.

Dataset Description:

Dataset files contain logs and production data from 2016-now, including power supply failures by substation and at barge, power consumption, current and production data as well as expert power database from 2019 including all grid interruptions data and extensive weather information.

The dataset is composed with real unfiltered data – since we would like to implement the model as a live model we want the model to be able filter "bad data" (Data containing words instead of values, for example: describing equipment communication error).

Projects Goals:

A major challenge in this project is to predict power supply failure using production indicators. We would like to receive a failure prediction at least 15 min in advance.

- **Basic:** anomaly detection in production and sensor data (high recall, low precision) - find a large volume of potential anomalies, which will then be analyzed and classified into critical/not critical by human experts, in order to improve recognition of critical failures.
- **Advanced:** Use basic goal results to create a predictive model with high precision (>98%) which can predict critical failures 15 min in advance. Model should detect >5% of critical anomalies before they occur.

Project Impact :

This project will have instant influence internally, as we will use it to predict and prevent power failures. This could prevent/shorten failure time resulting in less down time and less equipment = better production. Assuming this will give real value, it will be an important tool to the way we manage our energy system.

K Health

Domain:
Healthcare

Location:
Tel Aviv

Intro:

Healthcare systems are collapsing around the world. Half the world population has limited or no access to decent healthcare. In the west, healthcare costs are rising, waiting times are unbearable, and trust in healthcare and even in the science of medicine is at an all time low. Brilliant doctors are overworked, burdened with liability and outdated protocols, and forced to try to understand complicated medical conditions in just a few minutes.

K is building a solution. Over a million users in the US (#1 IOS health app) already talk with the K diagnostic bot for free, learning more about their potential medical conditions and making informed health decisions. Subscribed users can then chat with a live K doctor through the app, combining AI insights with a trained professional opinion. Soon, users will also be able to purchase medications, schedule physical appointments, and order lab tests seamlessly through the app, resulting in a quick resolution of health issues. And now, world-leading insurance companies are integrating K into their plans to offer more affordable and personalized care at scale.

Finally, with full access to our patients' medical history and outcome data, K will be effectively scaling clinical trials, learning faster than ever before how to provide the best health solutions for each person, understanding them as individuals and as a whole.

Logistics:

- Remote access to Maccabi servers.
- Preference to be onsite for at least a couple of hours every week.

#16

K Health

Type of task:
Clustering,
Information Extraction

Keywords:
NLP,
Ontology Learning
Domain Terminology Extraction,
Concept Discovery

From concept discovery to differential diagnosis

Problem Description:

Physician visit notes are one of the fundamental data types in electronic health records. These free text summaries of patient visits include most of the important information about the visits: medical history, symptoms, drugs, diagnosis, physical exams and more.

Through our partnership with Maccabi, we have access to over 400 million notes, collected over more than 20 years of medical practice, this data holds an infinite promise for groundbreaking medical research and AI solutions.

Information Extraction, the task of transforming raw unstructured text into a structured format, is a fundamental task in NLP. Transforming data into a rich ontology and a comprehensive knowledge graph, based on concepts and complex correlations found in the notes is one of the most important challenges in K Health, it is the key to doing analysis, building predictive models and finding exciting new medical discoveries.

This project touches on different aspects of NLP & ontology learning. This involves creating a data structure by identifying medical entities in textual data and mapping the free text visits into this structure.

Dataset Description:

- The dataset is free text summaries of medical visits.
- The dataset will be given in a CSV format.
- The notes are anonymized.

Project Goals:

- **Basic:** Domain terminology extraction & concept discovery. Extract general terms as well as terms associated with different medical conditions. Group terms into meaning bearing medical concepts, like different symptoms, physical exams, etc.
- **Advanced:** Task-oriented concepts. Build a diagnosis classifier (supervised) with your extracted concepts. Build an algorithm to identify which concepts are more important for the diagnosis of each medical condition.

Project Impact:

In this project you will be developing algorithms for expanding K's knowledge graph, and even build a knowledge graph from scratch. The knowledge base is at the core of K's research and any improvement to it will have both immediate and long-term impact on K's product.

Lightricks

Lightricks

Domain:
Creativity Apps

Location:
Jerusalem

Intro:

Lightricks is on a mission to create fun and powerful tools that reimagine the way content is created all over the world.

We empower aspiring artists, everyday individuals, and businesses with the best possible tools to unlock their creativity and create content that their audience connects with.

Whether you need to edit photos, video, sound, a portrait, a post or ad for social media - we have the most intuitive, brilliant tools for the best possible results.

ML Research at Lightricks:

Our products are built to be the best tools around, from design through engineering.

In order to allow for powerful capabilities and magical interactions, our ML research drives many of the features in our products, from automation of tedious tasks to enablement of complex creative processes for the laymen.

Logistics:

- Offices located in Jerusalem, in the Givat-Ram campus of the hebrew university.
- On-site presence required from time to time, most of the work can be done from home.
- Sign NDA, IP agreement (all IP produced will be owned by Lightricks).
- Work can be done on your own laptop, no sensitive data.

#17

Lightricks

Type of Task:
Generative

Keywords:
GAN,
Conditional Generation,
Neural Networks,
Style Transfer,
Harmonization

Multiscale style application for varied tasks

Problem Description:

Style application or transfer from one image to another is a well-researched problem.

Most solutions today focus on getting more of the original style or maintaining more of the details from the input image.

Our goal is to allow our users control over all these parameters at real time while also controlling more detailed parameters.

Such parameters may include the size of elements such as brush strokes, the amount of details transferred from each input and more complex compositions of the resulting image in different frequency bands of both the style and content images.

For that purpose, we propose a multiscale architecture that resembles some of the latest works in multiscale GANs alongside the conditioning of the result of such generative processes on a style image unknown at the time of training and provided at the time of inference.

Project Goals:

- **Basic goal:** Implement and train a multiscale conditional generation network architecture so that each scale stage is composed of a conditional style transfer network (e.g. WCT, MUNIT, etc.). Show improvement in this architecture's ability to maintain multi-resolution details compared to single-scale style transfer architectures.
- **Advanced goal:** Show use cases for this architecture in harmonization, super resolution and multiple instance generation.
- **Stretch goal:** Show control over scale-specific editing features of this architecture and/or submit to a tier-1 machine learning conference.

Project Impact:

Our Enlight suite of apps is always looking to improve the creative tools it provides for aspiring creators. Using this architecture will allow seamless composition of multiple images and changing the style of images in realistic ways while maintaining the full resolution details that our users expect us to preserve at all times.

LitiGate

Domain:
Legal tech

Location:
Tel Aviv

Intro:

LitiGate was founded in 2018 in the emerging field of legal-tech and selected as the most promising legal-tech startup in Israel for 2019.

LitiGate was created by litigators for litigators. It equips dispute resolution teams with an end-to-end operating system that automates repetitive tasks, facilitates real-time collaboration and uses cutting-edge AI to analyse evidence, assess risks and reveal untapped information from within case documents and relevant law.

We believe that machine learning, and deep learning in particular, will lead to a revolution in the day-to-day work of lawyers. As a leading legal-tech start-up, all the projects we propose are real-world open problems which we will tackle using innovative, state-of-the-art approaches.

Logistics:

Work location
Work will be from home. Guidance and advice will be given on a continuous basis.

Legal requirements
The students will need to sign an NDA (non-disclosure agreement) and IP waiver, as the projects includes access to internal Litigate knowledge, data and IP.

#18

LitiGate

Type of task:
Text Analysis,
Generative

Keywords:
Generation,
Paraphrase,
NLP

Paraphrase Generation

Problem Description:

Paraphrase generation is an on-the-rise topic. This task aims to generate a sentence, given a source sentence, by using different wording that conveys a similar meaning to the original sentence.

Paraphrases can be exploited in several ways. For instance, as a data augmentation module for training more robust classifiers. Another application is for sentence simplification in the case the original sentence is complex.

In recent years several approaches have been proposed, however, generating high-quality paraphrase is still considered to be a challenging task.

Dataset Description:

Dataset of several thousand source sentences and their corresponding paraphrases.

Project Goals:

- **Basic:** Develop an algorithm that will generate paraphrases given a source sentence and explore ways to incorporate the augmented data in the training of classification algorithm.
- **Advanced:** Develop a model that will learn to distinguish between a real and a generated sentence.

Project Impact:

We see a lot of value in this project as paraphrase generation can be used to enhance and make our existing algorithms more robust if used as a data augmentation module. It may also be used for sentence simplification purposes.

Matific

matific

Domain:
EdTech

Location:
Tel Aviv

Intro:

Matific is an adaptive, game-based, mathematics teaching and learning program designed to make elementary math learning fun! Matific is presently available in more than 60 countries and over 40 different languages. atific's pedagogy combines rich content with interactive games. Matific's content is carefully localised and mapped to each country's curriculum and popular textbooks. Matific is highly adaptive, allowing teachers to provide scaffolding and differentiation for each student. Matific users answer millions of questions daily across the globe. Matific empowers teachers and learners by presenting mathematics in a meaningful, comprehensive, and engaging way. Children learn at their own pace, improving their quantitative skills and analytic ability, receiving extra-practice in the form of gameplay.

Logistics:

Work location
Students work remotely using their own computer and are expected to visit us every week or two to discuss their progress.

51st Floor of Moshe Aviv Tower, Jabotinsky St 7, Ramat Gan
(8 min walk from Savidor train station).

Legal requirements:
NDA and agreement to release project outputs as MIT OSS must be signed by students.

Data access format:
Data will be provided in JSON format.

#19

Matific

Type of task:

Multi-label Classification

Keywords:

NLP,

Hierarchal classification,

Self-Supervised Learning

Curating the world's largest collection of interactive math problems

Problem Description:

Content creation is one of Matific's core competences, but being mostly performed manually it also expensive to scale. Licensing external content in electronic form somewhat alleviates the issue by switching from creation to curation. The first step in vendoring-in external content is assigning each item (problem) it's appropriate Matific-Curriculum entries – a multi-label classification problem, often performed manually and/or heuristically.

In this project you will combine several math Q&A datasets with Matific's expert-crafted content, effectively creating one of the world's largest collections of interactive curriculum-adjusted math problems.

Dataset Description:

Matific's collection of math (word-) problems, paired with correct answers and a multi-level (hierarchical-) index labelling each problem with one or more math-curriculum topics.

Several 100K, largely-unlabelled, answered math word-problems from partner organizations.

Project Goals:

- **Basic Goal.** The basic task is to correctly classify a given problem (-descriptor, from the labeled set) with the appropriate labels, at an increasing depth of the company's taxonomy tree starting at the top index level - comprising 10 labels such as "Algebra" or "Measurements", and increasing granularity up to over 350 leaf nodes (labels). Classification quality will be evaluated using a combined score based on both precision and recall, taking into account the structured nature of the hierarchical classification problem.
- **Intermediate Goal.** Inspect unlabelled problem collections and provide insights. For example, perform clustering to identify the set's own latent taxonomy (tree), map it to Matific's at various levels, and estimate taxonomy coverage and item distribution. Examine the validity of the Continuity-, Cluster-, and Manifold assumptions in preparation for the Advanced Goal step. Your success on this goal will be measured qualitatively by the breadth and depth of the insights provided.
- **Advanced Goal.** Apply your model to classify problems from our partner organization sets; e.g., employ self-supervised learning techniques, use Matific's true labels to classify increasingly larger subsets of unlabelled problem collections. A sample of true labels for used partner set problems will be obtained manually and serve to evaluate performance using the score established in the Basic Goal step.
- **Cherry-on-Top Goal.** Deploy your model(s) as a standalone web-service, implementing endpoints taking: a problem and level as input and returning the resulting labels [ranking over all nodes (possible labels) up to level, with confidence levels on each label, etc.]; or, given additional labels, return similar problems labelled with those. Your success on this goal will be measured qualitatively based on API legibility and code quality, and quantitatively based on latency.

Neurosteer

neurosteer mirror your mind

Domain:
Neuroscience

Location:
Herzliya

Intro:

Neurosteer's mission is to improve the quality of care of brain disorders and to enhance brain activity. For this, it develops a new generation of brain-sensing and control platform for a wide range of medical and wellness applications.

The cloud-based AI and Digital Signal Processing provides timely alerts and neurofeedback for optimal and personalized drug and neurostimulation intervention. Brain activity is being recorded by a proprietary sensor of a single EEG channel. This makes the continuous neurological sensor wearable, affordable and easy to use in the clinic and at home.

The current product provides 5 different cognitive biomarkers, as well as an indication of pre-ictal and epileptic seizure activity as well as sleep and consciousness levels. It is being validated in clinical trials in Israel, Europe and the US.

The company was co-founded by Prof. Nathan Intrator from the School of Computer Science and Neuroscience at Tel Aviv University. Prof. Intrator who is a world expert in machine learning and neural computation, has developed brain-inspired hybrid supervised and unsupervised architectures and addressed a variety of applications including blurred face recognition and multiple biosignal problems.

Logistics:

Work location
At least one day a week will be in the company's office in Herzliya. Monthly deliverables for the project will be defined and monitored.

Legal Requirements:
Legal requirements – IP ownership and confidentiality and HIPAA agreements required.

Data Access Format:
Unidentified subjects' data provided in CSV text format.
Data access format (own laptop/company hardware/publicly available etc.)

#20

Neurosteer

Type of Task:
Clustering

Keywords:
NLP,
Brain Computer Interface,
Sparse Clustering,
Research

Interpretation of Brain signals with Text Analysis Tools

Problem Description:

Brain activity can be viewed as a language with alphabet, words and grammatical rules. The goal of the project is to discover and assess the implied language of the brain and the complexity of that language under different brain conditions. Doing this will require to determine what is a robust brain alphabet and perform lexical analysis, using NLP tools assess the level of brain damage and create a BCI (Brain-Computer Interface).

Dataset Description:

Brain activity of healthy and sick individuals will be provided by Neurosteer. This will be in the form of vectors of 121 dimensions at every second. A couple of hundred hours of such data (from few hundred individuals) can be provided.

Project Goals:

- **Basic:** It is assumed that as the degree of brain damage of an individual is increased, the complexity of the “language” emanating from its brain activity is reduced. In this stage the main task is to define the “alphabet” of brain activity, from which grammatical rules can later be deduced.

  The initial goal is to create tools for automatic assessment of an individual with brain damage. This will rely on various clustering algorithms.

- **Advanced:** Following the basic stage, the next goal is to improve the interpretation of brain activity by applying the induced grammatical rules to create a more robust and comprehensive BCI. Specifically, given an alphabet found in the previous stage, use NLP tools to create a robust set of “words” and “sentences” from the alphabet. From that, one can infer some grammatical rules. For example, recognize which collections of symbols appear together with high probability

Project Impact:

Novel brain assessment tools, novel BCI for disabled subjects.

Neurosteer

Type of task:
Clustering,
Data Generation

Keywords:
Clustering,
Brain Computer Interface,
Data Generation,
Research

VR enhanced brain activity mapping

Project Description:

Construct a VR game with cognitive challenges and multiple actions (not motor but more cognitive). For example, it is known that Parkinson's and dementia, in general, reduce navigation abilities. A game that requires navigation (go back to fetch the key that was left in one room to open a box that is left in another room etc.) can challenge the memory and navigational skills of the elderly as well as retrain them to remember where they left the key etc.

The task is to use the actions performed during the game as labels to the brain activity during those actions and to create a model enabling some of the actions to be predicted by brain activity either completely, or probabilistically.

Tools to be used:

Clustering, deep learning and other ML models (Logistic Regression, SVM etc.) VR tools.

Data will be collected by the students using VR game.

Dataset Description:

Vectors of 121 dimensions that are created by Neurosteer brain interpretation.

Project Goals:

- **Basic:** Data generation - create automatic labeling of the data by interfacing the game (an appropriate open source game will need to be chosen)
- **Advanced:** machine learning using brain states as labels - create a model enabling some of the actions to be predicted by brain activity

Project Impact:

Novel brain activity modeling using VR, potential for novel BCI for disabled subjects.

Owlytics

OWLYTICS HEALTHCARE

Domain:
Healthcare

Location:
Tel Aviv

#21

Intro:

Owlytics was founded by professionals working in the healthcare information technology sector. Its principals are dedicated to improving the lives of seniors worldwide. The company is making a difference in the ongoing transformation of healthcare through the use of data technology.

Owlytics' system is a comprehensive solution aimed at providing seniors with safety, independence and better health. Designed to continuously and automatically detect falls, assess fall risk and provide early warning of potential health problems, it delivers actionable insights with proven outcomes that help provide better senior care while reducing cost

Logistics:

Work location
18th Floor of Atrium Tower, Located at Ze'ev Jabotinsky St 2, Ramat Gan.
8 minutes walk from Tel-Aviv central train station (Savidor).

Legal Requirements:
Sign NDA IP agreement (all IP produced will be owned by Owlytics)

Data Access Format:
Students work remotely using their own computer, and are expected to visit us every 2 weeks to discuss their progress.

Data will be provided in a csv or numpy array format that will be read and processed from company google drive using colab. Work will be done remotely from own laptop.

Owlytics

Type of task:

Classification and Segmentation of Time Series Data

Keywords:

Kinematic Analysis,
Health Monitoring,
Time Series,
Sensor Data

Elderly Human Daily Activity Recognition for Healthcare Using Wearables

Problem Description:

Having the ability to detect daily activities of seniors can help detect early signs of diseases, health issues, identify the risk of falling and offer pre-emptive treatment or rehabilitation for elders. In this project you will leverage multi sensor dataset collected form Owlytics users, to identify different daily activities performed by seniors such as sitting, standing, lying or walking.

Dataset Description:

The dataset includes real daily activities data on multiple sensors readings from smart watch and insoles.

The smart watch data consists of accelerometer and heart rate and insoles pressure sensors from heel and front of the foot, gyroscope and accelerometer

Project Goals:

- **Basic Goal:**
  The basic task is to classify the type of activity being performed to one of the following classes: lying, sitting, standing or walking.

  Classification quality will be evaluated using a combined score based on both precision and recall.
- **Advanced Goal:**
  Pointwise segmentation of walking. This is the equivalent of semantic segmentation of an image, in which we are classifying each pixel of an image to one of a given set of classes. In this time series task, you will classify each point in time to either walking or not.

Project Impact:

Being able to classify the activity of elderly persons is a fundamental task for automatic health assessment and fall risk detection. A successful classification and segmentation would allow better understanding and evaluation of fall-risk and prevention of health deterioration.

SimilarWeb

SimilarWeb

Domain:
Market intelligence

Location:
Tel Aviv

Intro:

SimilarWeb is a world leader providing insights on more than 80M websites across 200+ industries and +190 countries. Using Machine Learning (ML), our core product relies on transforming various raw data sources of online activity into traffic estimations (e.g. estimated number of visitors to the site, number of sessions, average time on site, etc). SimilarWeb's estimation data are widely used by various clients, spanning from SEO experts, sales professionals, retailers, private and public investors and many more to better understand global market trends of the digital world. We process over 3 billion raw events every day and apply ML to solve various probelms:

- Correcting biased samples using data from direct measurements
- Transfer learning from one domain to another
- Fusion of multiple data sources into stable traffic estimations
- Time series problems
- Classification of traffic sources
- Graph classification and embedding
- Optimization at scale

Logistics:

In the projects we suggest, you will help DS and engineering teams working on developing our core products and features. We will meet once a week at our offices to a project sync, and we also encourage you to work at our offices once a week. At the end of the project, you will present your work to our teams and provide us with a technical report and your code (we will open together a git repository).

#22

SimilarWeb

Type of task:
Prediction,
Supervised Learning

Keywords:
Web Traffic,
Embedding,
Text Processing,
Social Data

Website Traffic Prediction with Twitter Data

Problem Description:

Obligated to the quality and accuracy of our product, we constantly enrich our traffic estimation pipelines with new data sources. One of the most interesting publicly - available datasets in the world is from Twitter. It's clear that many sites use social networks, including Twitter, to drive traffic into their websites, and there exists a strong correlation between website traffic to its Twitter presence. Our goal in this project is to develop a traffic estimation pipeline for top 10K sites in the US using Twitter data.

Dataset Description:

- **Labels:** SimilarWeb traffic estimations per website (number of visits - unique users per day)
- **Input to features:** Twitter Data - will have to be scraped by the project team during the project - many websites own a twitter account, where they publish links to their websites (usually news, media and e-commerce). You will query Twitter's API to extract all relevant tweets that include a link to the relevant website and all the metadata included in that tweet (likes, re-tweets, comments, users etc). You will use this information to build your features and embeddings to be used in your traffic estimation model.

Project Goals:

- **Basic:** Estimate (predict) US top 10k websites traffic using Twitter Data.
- **Advanced:** Use other publicly available social APIs to improve previously obtained results of estimate traffic, e.g. Facebook, reddit, Quora.

Project Impact:

Each estimation pipeline we add to our heavy estimation machinery helps us to enrich and improve the robustness of our platform and the reliability of our data for our clients.

SimilarWeb

Type of task:
Prediction,
Supervised Learning

Keywords:
NLP,
Sentiment Analysis,
Web Traffic,
Search Data,
Stock Price Prediction

Stock price prediction using keyword data

Problem Description:

Many investment companies are interested in predicting market prices using alternative data; for companies with a large digital presence, SimilarWeb data provides a proxy for a company's performance by estimating the company's online web activity. The hypothesis we want to validate is whether search queries that relate to a given company can correlate to its stock performance. For example, using sentiment analysis, we can see whether the overall sentiment of the keyword dataset contributes to the price fluctuations of a stock.

Dataset Description:

- Search queries that lead to a website via a search engine with its related estimated traffic in 2019 (e.g. the query "Black Mirror Netflix" lead to 10K visits to Netflix.com in 2019), in any and all countries.
- Publicly available stock prices of a company (daily, monthly).

Project Goals:

- **Basic:** Build a model that predicts stock performance (direction or absolute value) based on search queries for that website using information in SimilarWeb: estimated traffic, other websites this query lead to.
- **Advanced:** Enhance previous model with more information that can be extracted from search queries and may correlate with stock performance (e.g. topic, industry category, intent, sentiment and major word classes such as verb, noun etc).

Project Impact:

Being able to predict stock performance will allow SimilarWeb's investor clients to consider SimilarWeb as their digital data provider of choice when making investment decisions. It also allows us to understand how public perception and user engagement can affect stock performance.

SodaStream

Domain:
Manufacturing

Location:
South Region

#23

Intro:

SodaStream develops, manufactures, and markets home sparkling water makers and related products. SodaStream's sparkling water makers enable consumers to transform ordinary tap water into sparkling water and flavored sparkling water at the touch of a button. Products are available at more than 80,000 retail stores across 45 countries. SodaStream products are environmentally friendly. One SodaStream reusable bottle rids the world of up to 3 000 single-use plastic bottles.

SodaStream's CO2 refills is one of its main growth pillars. The company operates in a razor and a razor-blade model. The razor is obviously the soda maker and the CO2 refills are the main blade (in addition to the flavor syrups and the bottles). It has a significant future revenue stream. SodaStream is selling over 40M CO2 refills a year. The company operates 6 filling stations around the world.

ML in the Company:

As part of the company's evolvement, SodaStream has decided to invest resources and develop the fields of Machine Learning and Data Science, although we currently don't have any expertise in this area. We are now building the required infrastructure for it. The acquisition of SodaStream by PepsiCo opens new opportunities to work together in these fields and we have already started exploring different cooperation possibilities.

Logistics:

Work location – Kfar Saba headquarters office / Lehavim factory (Gas fill station). It is possible to work from home.

SodaStream

Type of task:

Predictive Maintenance Analytics,

Unsupervised Machine Learning

Keywords:

Predictive Maintenance Analytics,

Sensor Data,

Production Line

Forecast of production-line delays in CO_2 filling lines

Problem Description:

SodaStream manages multiple gas filling stations around the world. Every filling line consists several filling heads. An issue or delay in a single filling head can influence the pace of the entire line. The filling lines are automatic and have sensors monitoring parameters in every filling head of every line.

Today, we know how to react to problems on the line as they arise, but we are looking for means to predict problems based on our previous data, allowing us to respond before or very soon after they arise and before significant delays and damages are caused. We are also looking to find the best combination of parameters that will allow us to maximize the outputs.

Dataset Description:

Data is well structured contains archive tags for all sensors in Israel and international filling lines. It includes a large range of parameters including filling head speed, CO2 weight in each cylinder being filled, gas density, gas temperature, etc.

Project Goals:

- **Basic goal:** Analyze and understand the correlation between measured values from sensor data to production outputs. Perform analysis and feature engineering in order to locate factors correlated to changes in production outputs.
- **Advanced goal :** Create a predictive model in order to alert for an expected decrease in production rate based on a change in the sensor values measured on the line and the correlations found in the basic goal.

Project Impact:

Help the company to improve outputs of cylinder refills and increase quality level by preventing production delays or creating an early warning allowing for faster response time in fixing such delays if they arise. As the CO2 refills take a major role in the company’s business model, every improvement in the outputs can make a big difference.

YANDEX SCHOOL OF DATA SCIENCE

Trigo

trigo

Domain:
Retail tech

Location:
Tel Aviv

Intro:

Founded in 2018, Trigo empowers grocery retailers to seamlessly eliminate the industry's leading consumer pain point -- checkout -- through a world-leading computer vision and AI-based system.

Developed by a team of leading researchers from academic, industry and military backgrounds, our solution is based on ceiling-mounted commodity cameras, proprietary algorithms and neural networks. The system tracks the movement of shoppers and goods throughout the store, automatically tabulating each shopper's total via a mobile app.

Trigo has over 50 employees, with offices in Tel Aviv and London.

We are working with a number of global grocery chains, including Tesco, the fifth-largest grocery retailer in the world. Our system is already deployed in stores as large as 5,000 square feet, covering thousands of products in each store.

Machine Learning at Trigo:

Trigo has a strong team of scientists and research engineers working on a wide variety of deep learning, computer vision and machine learning algorithms. The team deals with hard problems, such as pose estimation and tracking, object detection and tracking and fine-grained classification, which are essential in building an autonomous store. Therefore, machine learning and specifically deep learning are at the core of Trigo’s business.

Logistics:

Work location:
Work on the project can be conducted remotely, with periodic F2F meetings.

Legal requirements:
Sign NDA, IP agreement

Data access:
Students will be expected to have their own laptops.

#24

Yandex

Trigo

Type of task:

Pose Estimation

Keywords:

Self-Supervision,

Semi Supervised Learning,

CNN,

Deep Learning

Self-supervised Multi-Person Pose Estimation in Crowded Scenes

Problem Description:

Estimating the pose of multiple people in images is a fundamental computer vision task, which has attracted tremendous interest for its numerous applications. Recent advances in human pose estimation have been achieved by harnessing the power of deep convolutional neural networks and large-scale pose estimation datasets.

However, correctly applying pose estimation to real-world problems requires massive amounts of well-curated data. This is a notoriously challenging task, hindered by privacy issues and annotation costs. Therefore, self-supervised and semi-supervised pose estimation algorithms are of prime importance to leverage the vast amount of unlabeled data available.

Applying pose estimation to real-world applications presents a wide array of challenges that are rarely addressed in the literature. One such challenge is crowded scenes, which introduce difficulties such as people in close proximity, mutual occlusions, and hindered or partial visibility.

Self and semi-supervision methods have traditionally been applied to classic computer vision problems such as classification. In this project, you will take a novel approach, applying these methods to the task of pose estimation. You will be given a set of deep neural network architectures and training methods and will be tasked with improving the resulting models using additional unlabeled data.

Project Goals:

- **Basic goal:** Use semi-supervision methods and convolutional neural networks (CNN) to harness unlabeled multi-person images towards improving pose estimation with a limited amount of labeled data.
- **Advanced goal (1):** Improve model performance specifically on challenging inputs such as crowded scenes.
- **Advanced goal (2):** Further use temporal information and combine self-supervision with time-series analysis on videos.

Project Impact:

Accurately predicting human poses is crucial for a smoothly functioning autonomous store. However, annotating data for pose estimation is an extremely expensive endeavor. Thus, improving algorithms by utilizing unused, unlabeled data is of great importance and can make a significant difference, especially if it can be applied in crowded environments.

Valerann

VALERANN

Domain:
Smart Transportation

Location:
Tel Aviv

Intro:

Valerann is a leading data provider helping to accelerate the revolution of the intelligent transportation industry. This is made possible through the "Smart Road System", an end-to-end solution integrating innovative sensing technologies into roads, transforming them into a future-ready, connected infrastructure that provides a comprehensive new source of data.

Valerann solves this problem with its innovative Smart Road System (SRS), an end-to-end road management and data solution. We gather granular, high-resolution data every 10 meters along the road. We aggregate and analyze it in real time. And we share critical, actionable intelligence: identifying safety risks such as accidents, stopped vehicles, or debris; predicting worsening traffic congestion by measuring lane-by-lane traffic flows; and monitoring weather conditions to recommend road servicing.

Valerann's wireless sensors are multi-sensory cross-signal units, that are implemented within the road, collecting data from their surroundings and send it in real-time to Valerann's gateway. They are solar-powered, include a proprietary wireless communication system and active RGB illumination.

The Valerann cloud combines the information collected from all gateways and use machine- learning algorithms to map, track and predict everything that takes place on the road. This results in a comprehensive, high-resolution platform of real-time data.

Logistics:

- Work Location: One day attendance at our office. We will offer consultation, answer questions and give advice during this day
- Need to sign an NDA and IP documents

#25

Valerann

Type of task:
Supervised-Learning

Keywords:
IoT,
ITS,
image Segmentation

Vehicle in-lane position prediction using magnetic sensor images

Problem Description:

Valerann has a smart road system that is installed in many highways around the world. This IoT solution produces different kinds of data, one of them containing magnetic data signatures of potential passing vehicles. In this project you will be asked to create a model that given an array of sensor data predicts the position of the corresponding vehicle.

Dataset Description:

We will provide you with access to our sensor datasets and API images dataset.

The images dataset consists of images and corresponding in-lane position (x,y) and event IDs, including human-labelled images which can be used as the labels for the sensors data target values.

Magnetic signature dataset – this is the sensors data, which consists of the raw magnetic event sensed by the sensor for the corresponding event's ID.

Since both datatypes represent the same reality, we want to find the correlation between the two, to better predict vehicle position.

Project Goals:

- **Basic Goal:** Create a model that given an array of magnetic events, predicts in-lane position
- **Advanced Goal:** Create a model that predicts the distance between the given vehicle's wheels

Project Impact:

If successful, your model will be used to predict in-lane position of vehicle and will greatly help location adjustments done by autonomous vehicles and connected cars

WeissBeerger

WeissBeerger

A part of the ABInBev Family

Domain:
Retail tech

Location:
Tel Aviv

Intro:

WeissBeerger is connecting the on-premise outlets such as bars, restaurants, cafes by getting access to their transactional data from their POS machine. The data is unstructured and not standardize since most of the products does not carry barcode, a glass of beer ordered in the bar is not barcoded, and bar owners can type in their POS machine random text that represent the transaction.

For instance: transaction consist a 500ml glass of Stella Artois can be reflected as:

- Bar 1: STL big
- Bar 2: Stella big glass
- Bar 3: Stella 500ml

In Weissbeerger system it should have the same SKU ID, therefore we developed an automated "matching" algorithm backed by humans, this algorithm will take the text, extract values using NLP and classify into attributes using machine learning.

Once the data is standardized in our system, we can start creating value. The data is being used to understand consumer behavior in the on-premise channel and find what make their decisions during their visit in the bar and how can we impact bar performance and category (beer) performance in the channel. Few examples of initiative we are solving:

1. Predicting patterns ahead the curve: utilize transactional data and social media data to predict virality of new brands / beer types in the market based on time series analysis
2. Recognize consumer journey – understand in what order consumers consume products in bars based on labeled data. Cluster type different consumers journeys to meaningful business clusters
3. Bars survival – business like bars survive for 28 months on average before getting into financial problems. We are running a prediction model to recognize indicators for "dying" bars based on feature importance process and trying to predict which bars will have low survivor rate.

Logistics:

- Work location: on-site presence required, students can complete work from home
- Legal requirements: sign NDA

#26

Yandex

WeissBeerger

Type of task:
Prediction

Keywords:
Time Series Analysis,
Pattern Recognition,
Feature Importance

Prediction ahead of the curve

Problem Description:

We want to predict patterns of new brands or new beer types becoming viral ahead the curve. We can utilize transactional data and social media data to predict virality: sales curve, social media semantics, patterns in check and more.

Dataset Description:

Transactional data include all orders to consumers in bars, each order have information of all items in orders, the categories, quantity and price. Meta data on bars, and products, no social media data. Rest of the information can be gained by extensive data exploration.

Projects Goals:

- **Basic:** Predict beer consumption by brand per location based on historical demand
- **Advanced** – Use basic prediction results and feature importance mapping to attempt to predict future trends and changes in consumption, such as when a beer brand or type may become viral and grow significantly, and predict trend behavior - when it will converge / stop / regress.

Project Impact:

Success in this project will allow us to add the new capability to Trayz (our product for business).

Yandex Zen

Domain:
Content discovery

Location:
Moscow

Intro:

Yandex.Zen is an algorithmic feed that matches millions of users with engaging content to read and watch.

We've got our own publishing platform and now have over 30K active publishers covering a wide range of user interests.

We've got 14M daily active users and they spend 40 minutes daily on average in our feed.

ML in Yandex.Zen:

At Yandex.Zen we have to rank millions of web pages in a fraction of a second for every user request. We employ a wide range of ML algorithms like matrix factorizations, neural networks and gradient boosting to name a few. We've got 10K user requests per second, so we have to write a highly optimized code that employs vectorized instructions, KNN indices and cascades of ranking algorithms to squeeze in 500 ms per request. We also train a lot of classifiers to fight clickbait, categorize articles, fight fraud, etc.

Logistics:

- Work location: remote work from home, weekly consultations in Skype
- Legal requirements: signing NDA, IP agreement (all IP produced will be owned by Yandex)
- Data access: dataset is hosted on Yandex.Disk, all work can be done on your local computer or in the cloud.

#27

Yandex Zen

Type of Task:
Recommender Systems

Keywords:
Ranking,
Matrix Factorization,
Neural Networks,
NLP,
Web Traffic Data

Improving cold start in an algorithmic newsfeed

Problem Description:

At Yandex.Zen we recommend millions of unique articles each day. We're constantly improving our recommender engine to find the most suitable content for our users.

Some of our users interact with the feed for the very first time and we want to make that experience as personal as possible. Because Yandex is a search engine, we know something about the prior user activity on the Internet, like what pages the user visited. We're seeking ways for better utilization of that information.

It will be your task to come up with the algorithms that can better match the external user history and clicks that he/she makes in our feed. We believe there's a correlation, and we need the best models to unveil it.

Dataset Description:

We will provide the external history for a million of users coupled with their events in our feed.

External events include only clicks on web pages. But feed events include both clicks and show (or impression) events for article web pages. To better protect the privacy of our users, for each web page you will have a hash of the url, a hash of the page domain (like cnn.com) and hashes of page title words.

Project Goals:

In this project, you will strive to predict which pages the user will click. We will measure the success in terms of NDCG ranking metric, which tends to correlate with our online performance.

- **Basic:** Apply classic recommender system techniques for click prediction.
- **Advanced:** Build up a custom model which utilizes both text and collaborative information to analyze the implicit external user history.

Project Impact:

Successful implementation of this project will help us to better jump start recommendations for users with external history (roughly 50% of all new users) and improve user retention and other business KPIs.

YANDEX SCHOOL OF DATA SCIENCE

Yandex