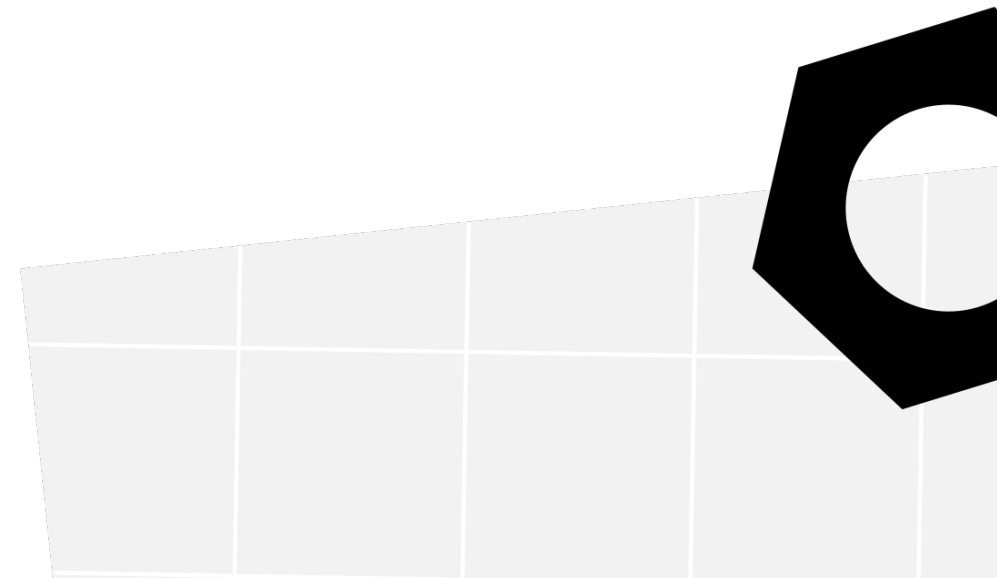




Projekt końcowy

Kurs Data Science





Założenia projektu końcowego:

- Praca na platformie GitLab
- Grupy maksymalnie 3 osobowe (może być mniej osób)
- Projekt musi być zgodny z zasadą SMART
- Do dyspozycji są wstępnie opracowane tematy, ale obowiązuje pełna dowolność
- Projekt to nie tylko model, ale też jakaś namiastka finalnej aplikacji (może być nawet aplikacja komunikująca się z użytkownikiem w konsoli)
- Projekt w Pycharmie, jeżeli jest taka potrzeba należy przenieść potrzebną część kodu na dysk i skorzystać z collaba



Plan prac:

Tydzień 1

1. Warsztaty z Design Thinking. Następnie określenie zarysu projektu (1 dzień)
 - Do zrobienia na sesji: wybór tematu, wybór i wstępna analiza danych, uzasadnienie, prezentacja na ten temat
2. Na początku prezentacja grup – jaki temat wybrali, jakie ma uzasadnienie społeczne, biznesowe, jakie zbiory danych (2 dzień)
 - Do zrobienia na sesji: poprawki po konsultacji z trenerem, dogłębna analiza danych, wybór modeli, uczenie, pierwsze wyniki



Plan prac:

Tydzień 2

1. Prezentacja postępów, jakie modele zostały wybrane, jakie są pierwsze wyniki, co sprawiło problem, co można poprawić (3 dzień)
 - Do zrobienia na sesji: poprawki po konsultacji z trenerem, badania modeli
2. Prezentacje końcowe, wyniki – prezentacja techniczna i **pitch-deck** (4 dzień)
 - Do zrobienia na sesji: opracowanie wyników, prezentacja końcowa

<https://marketingibiznes.pl/start-up-zone/pitch-deck/>

Design thinking



Design thinking - co to jest?

Design Thinking to podejście do tworzenia nowych produktów i usług w oparciu o głębokie zrozumienie problemów i potrzeb użytkowników.





Design thinking – co to jest?



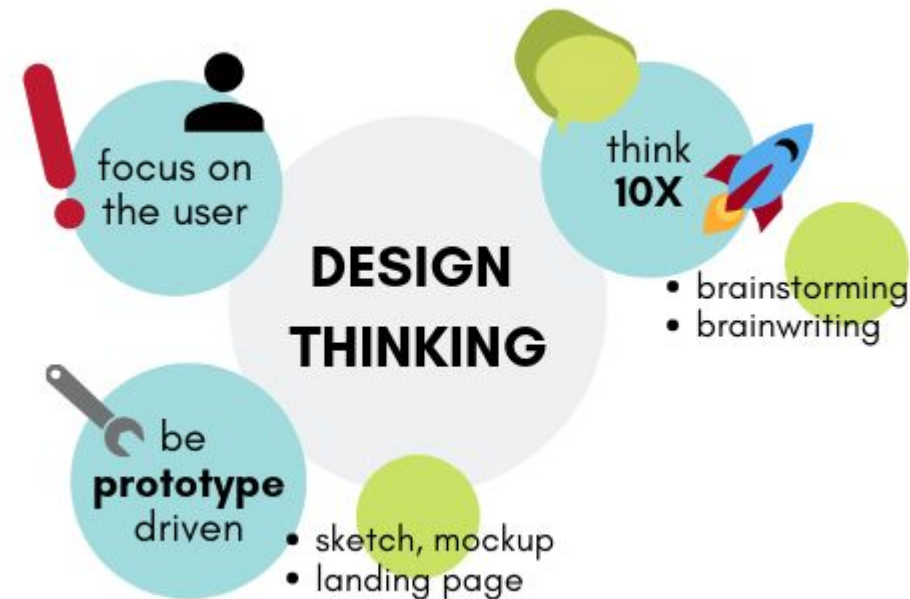
Design Thinking działa w oparciu o następujące założenia:

1. **Koncentracja na użytkowniku** – dogłębne zrozumienie jego uświadomionych i nieuświadomionych potrzeb.
2. **Kreatywna kolaboracja** – spojrzenie na problem z wielu perspektyw, szukanie nowych rozwiązań, wyjście poza utarte schematy.
3. **Eksperymentowanie i testowanie hipotez** – budowanie prototypów i częste zbieranie informacji zwrotnej od użytkowników.

Design thinking – co to jest?

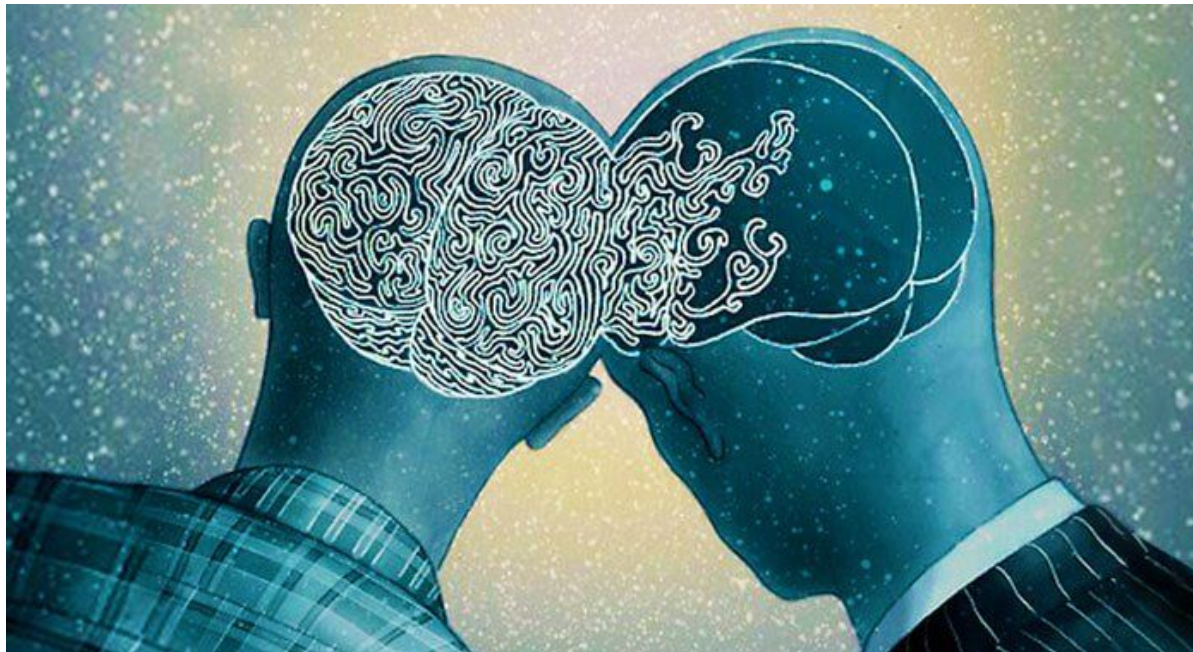
Celem projektów realizowanych zgodnie z Design Thinking są więc rozwiązania:

- Pożądane przez użytkowników
- Technologicznie wykonalne
- Ekonomicznie opłacalne



Faza empatyzacji

Proces Design Thinking zaczyna się od empatii. Pierwszym etapem jest głębokie zrozumienie potrzeb i problemów użytkownika. Kluczowe jest rozpoznanie problemów i motywacji, które mają wpływ na ludzkie wybory i zachowania.





Faza empatyzacji

W tym celu używa się takich narzędzi jak:

- mapy empatii
- wywiady etnograficzne
- obserwacje użytkowników
- ankiety rozpoznawcze wraz z dokładną analizą środowiska (hit the streets) i potrzeb w kontekście funkcjonalności.

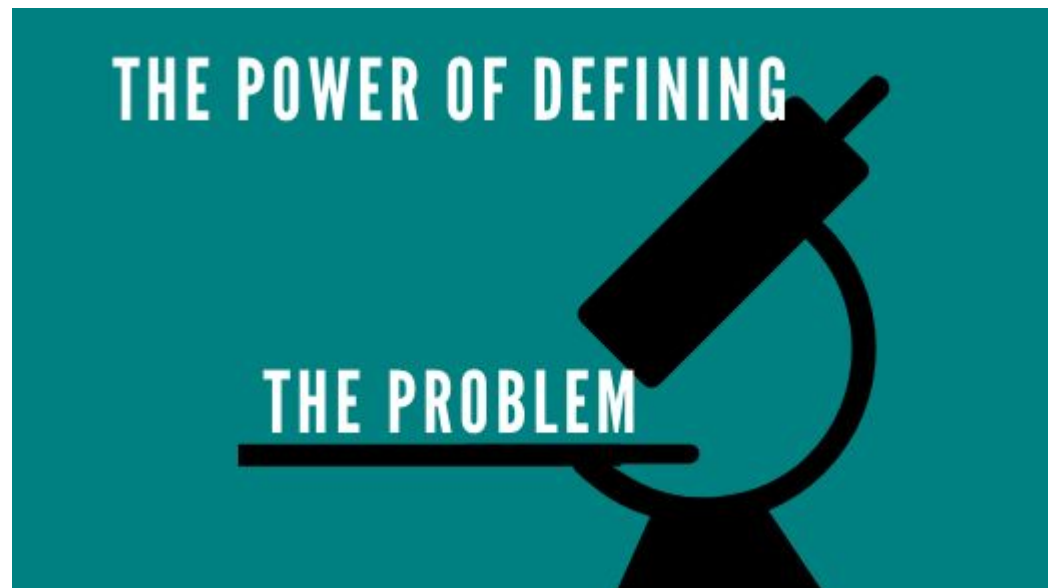
Należy przy tym zwrócić uwagę, iż tradycyjny focus group się tu nie sprawdza, ponieważ ludzie mają tendencję do racjonalizowania swoich wypowiedzi i unikania komentarzy krytycznych.

Dyskretna obserwacja zachowania może wykazać, że użytkownicy stosują jakieś własne amatorskie usprawnienia, które mogą stać się inspiracją dla nowych produktów.



Faza definicji

Na tym etapie zespół dokonuje syntezy informacji zebranych podczas Fazy Empatii w celu zdefiniowania co jest właściwym problemem. Etap ten wymaga przełamania ram myślowych i przyzwyczajień, które ograniczają pole widzenia.



Faza definicji

Z(re)definiowanie problemu może diametralnie zmienić kierunek poszukiwanych rozwiązań. Etap ten bywa olbrzymim wyzwaniem, ponieważ większość osób woli od razu pracować nad konkretnym rozwiązaniem, a nie poruszać się w niepewności wielu możliwych kierunków. Zbyt szybkie zdefiniowanie problem zawęży pełny obraz. Może się okazać, iż zainwestowane pieniądze, czas i energia nie adresują właściwego problemu. Porażka!

Przy definiowaniu problemu można się wspomagać takimi technikami jak technika 5x why, mapowanie problemu na osi **jak? vs po co?**





Faza ideacji

Na tym etapie zespół koncentruje się na wygenerowaniu jak największej ilości możliwych rozwiązań dla zdefiniowanego problemu. Wymaga to nie tylko silnego zaplecza merytorycznego, ale przede wszystkim odwagi w kreowaniu nowych, nieszablonowych rozwiązań oraz powstrzymywania oceny i krytyki pomysłów pozostałych członków zespołu. Etap powinien zakończyć się oceną i demokratycznym wyborem najlepszego pomysłu, na bazie którego powstanie prototyp.

Podstawowym narzędziem jest tutaj proces burzy mózgów – Brainstorming. Należy pamiętać, iż burza mózgów nie jest celem samym w sobie, a jedynie punktem wyjścia do określenia kolejnych kierunków działań. Głównym wyzwaniem jest przestrzeganie podstawowych zasad brainstormingu: proponuj nawet najbardziej szalone rozwiązania, nie oceniaj, buduj na pomysłach innych, nie przyzwyczajaj się do swojego pomysłu, pozbądź się ego, nie koncentruj się na ograniczeniach. Przysłowiowe “kolorowe karteczki na ścianie” służą do zmapowania procesu myślowego, są tymczasowe, można je swobodnie przeklejać, układać w różnych konfiguracjach przez co przypominają, że proces wymaga sporej elastyczności i dystansu do własnych pomysłów.



GENEROWANIE
POMYSŁÓW



Faza budowania prototypów

Na tym etapie powstaje fizyczny prototyp, ale celem nie jest tworzenie skomplikowanych modeli o cechach zbliżonych do produktu końcowego. Najważniejsza jest możliwość wizualnego zaprezentowania pomysłu użytkownikom i szybkie zebranie opinii na temat rozwiązania. Lepiej szybko dowiedzieć się, że nasz pomysł rozmija się z realnymi potrzebami i zmienić kierunek niż brnąć w kosztowną budowę czegoś wg. naszej własnej „idei fixe”. Nigdy nie można mieć pewności, że produkt końcowy będzie sukcesem, ale częste budowanie udoskonalonych prototypów, oddawanie ich w ręce użytkowników i słuchanie co mają do powiedzenia zmniejsza ryzyko końcowej porażki.



Faza budowania prototypów

Do budowania szybkich prototypów można użyć kartonu, drewna, styropianu... właściwie czegokolwiek. Prototyp nie zawsze musi być przedmiotem – w przypadku usług można się posłużyć storyboardem czy rysunkiem ścieżki użytkownika. Ważne, żeby zrobić krok dalej niż słowny opis i w dowolny sposób zwizualizować nasz pomysł.





Faza testowania

Na tym etapie wybrane rozwiązanie jest testowane w środowisku użytkownika. Ważne jest przede wszystkim określenie parametrów koniecznych do spełnienia tak, aby jednoznacznie określić wynik przeprowadzonego testu. Etap ten wymaga zaangażowania wielu stron i wsparcia od strony technicznej, formalnej, administracyjnej, prawnej.

A code tester walks into a bar.
Orders a beer.
Orders ten beers.
Orders 2.15 billion beers.
Orders -1 beers.
Orders a nothing.
Orders a cat.
Tries to leave without paying.



Faza testowania

Ważne, aby proces testowania odbył się w realnym środowisku, w którym produkt będzie używany. Dopiero po testach zakończonych pozytywnym wynikiem możemy mówić o tym, iż dany produkt, usługa są gotowe do ostatecznego wdrożenia. Niestety etap testowania jest często pomijany przy realizacji wielu przedsięwzięć co powoduje, iż z pozoru dobre pomysły i rozwiązania są bezpośrednio implementowane do codziennego użytku i dopiero tam okazuje się, iż nie spełniają one wymaganych założeń i oczekiwań odbiorców.





Polecany materiał:

- <https://dschool.stanford.edu/resources/the-bootcamp-bootleg?fbclid=IwAR0rK7TTgMsnXd4GXL2RupUB5gkPXsZwq2MYflyc3oVVxP-KKoJs8ls8FAs>
- Świetny bootcamp w formie książki o Design Thinking (naprawdę polecamy, rozbudza kreatywność :)



Spróbujmy!



- Podzielmy się na dwuosobowe grupy. Niech każda osoba z pary będzie jednocześnie wykonawcą i odbiorcą.
- Dla ułatwienia komunikowania można skorzystać z tego komunikatora:
<https://gather.town/app/6Q95DO7QqEWxg26R/sda>
(przy danym biurku siedzą osoby z grupy, wtedy słyszą tylko siebie)
- Naszym zadaniem będzie stworzenie idealnej, brakującej aplikacji mobilnej dla drugiej osoby z naszej grupy.
- Na początku przeprowadźmy fazę empatyzacji, porozmawiajcie o Waszych potrzebach, czego Wam brakuje w obecnych aplikacjach (może to być coś totalnie kosmicznego ;))
- Następnie niech każdy z Was przejdzie do fazy definicji oraz ideacji. Po każdej z faz przedyskutujcie ze sobą Wasze poczynania, aby odrzucić nietrafione pomysły.
- Potem przystąpcie do tworzenia prototypów. Mogą to być ekrany aplikacji narysowane w Paincie, mogą być grafiki przedstawiające schemat zdarzeń. Pobudźcie wyobraźnię.
- Na koniec przejdźcie do fazy testów i zapytajcie odbiorcę, jak mu się podoba prototyp. Czy jest potrzebna iteracja?
- Czas na wykonanie ćwiczenia: **60 minut!**

Omówienie proponowanych tematów





Propozycje projektów:

- Analiza sentymentu wypowiedzi polityków z Twittera (dane z mediów społecznościowych)
- Autonomiczny pojazd (przetwarzanie obrazów)
- SMOG (wiele danych z różnych źródeł)
 - <https://www.kaggle.com/datascienceairly/air-quality-data-from-extensive-network-of-sensors>
- Predykcja opóźnień w liniach lotniczych (wiele danych z różnych źródeł)
 - <https://www.kaggle.com/usdot/flight-delays>

Propozycja temat 1: Analiza sentymentu wypowiedzi polityków





Przedstawienie problemu



- Politycy w Polsce (na świecie również) posługują się często ostrym językiem, mocno nacechowanym emocjonalnie
- Powoduje to duże spolaryzowanie społeczeństwa, które powinno razem dążyć ku lepszej przyszłości
- Zadanie ma na celu przeanalizować wypowiedzi polityków kilku najpopularniejszych partii w Polsce i zbadać ich nacechowanie emocjonalne
- Sprawdźmy, która partia wydziela najwięcej hejtu!



Zakres/wyzwania



- zrozumienie i pozyskanie danych z Twittera/innych portali
- podstawowa analiza językowa
- wizualizacja serii czasowych
- notatniki Jupyter
- python:
 - przetwarzanie danych: pandas
 - grafy: networkx, igraph
 - wizualizacja: matplotlib, seaborn, bokeh, ggplot, wordcloud, folium
 - uczenie maszynowe: scikit-learn, xgboost
 - NLP: gensim, spaCy, tweet2vec, polyglot
 - statystyka, algebra: numpy, scipy
 - scrapping danych: newspaper, portia, pyspider, scrapy, twint
- praca zespołowa



Dane



Dane musimy pobrać z serwisów informacyjnych (głównie z Twittera, ze względu na frameworki umożliwiające scrapping, ale jeśli komuś się uda również Facebook, Instagram, może jakieś wywiady).

Powinniśmy zebrać wypowiedzi kilku polityków z każdej z największych partii w Polsce (co najmniej 5) + reakcje na te wypowiedzi (sprawdzenie, co się ludziom podoba).

Zbiór danych powinien mieć co najmniej 5 kolumn -> dane polityka, data wiadomości, treść, liczba polubień, liczba komentarzy.





Analizy i wizualizacje



- A. Pokaż liczbę tweetów w czasie (oś x) dla kandydatów za pomocą dowolnego wykresu liniowego; pogrupuj dane dla partii.
- B. Pokaż popularność tweetów w czasie (liczba polubień, liczba retweetów) – polubienia przypisane do daty publikacji tweeta.
- C. Pokaż rozkład aktywności w zakresie dnia (w ujęciu godzinowym) i tygodnia (w ujęciu dziennym) dla aktywności tweetowania.



Analiza sentymentu

Używając odpowiedniego modułu Pythona pokaż wizualizację sentymentu tweetów (średnia) w czasie dla wszystkich polityków. Pogrupuj wypowiedzi względem partii. Wyciągnij wnioski. Zobacz jak ludzie reagują na tweety o danym nacechowaniu.



Pomysły do dalszego rozwoju projektu:

- Pozyskanie informacji o retweetach
- Analiza danego tematu, np. lockdownu i informacji z nim związanych
- Pozyskanie informacji z innych lat, innych polityków, innych krajów

Propozycja tematu 2: Pojazd autonomiczny





Przedstawienie problemu




- Pojazdy autonomiczne zyskują coraz większą popularność i możliwe że w najbliższym czasie w pełni zastąpią manualny sposób prowadzenia pojazdów
- Algorytmy, które kierują takim pojazdem, działają między innymi na podstawie systemów wizyjnych
- Zadaniem jest stworzenie algorytmu rozpoznającego znaki pionowe oraz pozycję pojazdu względem pasa ruchu



Zakres/wyzwania



- wyszukanie i pozyskanie danych zawierających obrazy drogowe (zanotowane ze znakami i liniami)
 - wstępne przetwarzania obrazu
 - zastosowanie sieci neuronowych
 - notatniki Jupyter
 - python:
 - przetwarzanie obrazu: opencv
 - sieci neuronowe: tensorflow, pytorch
 - metody konwencjonalne: progowanie, detekcja krawędzi Canny'ego, klasyfikatory HOG
 - uczenie maszynowe: scikit-learn, xgboost
 - tracking detekcji
 - praca zespołowa
- 



Dane



Dane trzeba podzielić na wykrywające znaki pionowe:

- <https://benchmark.ini.rub.de> - zbiór niemieckich znaków
- <https://www.kaggle.com/c/traffic-sign-recognition/data> - różne znaki w konkursie
- <https://www.cvl.isy.liu.se/research/datasets/traffic-signs-dataset/> znaki Szwedzkie

oraz takie wykrywające linie (znaki poziome):

- http://www.cvlibs.net/datasets/kitti/raw_data.php



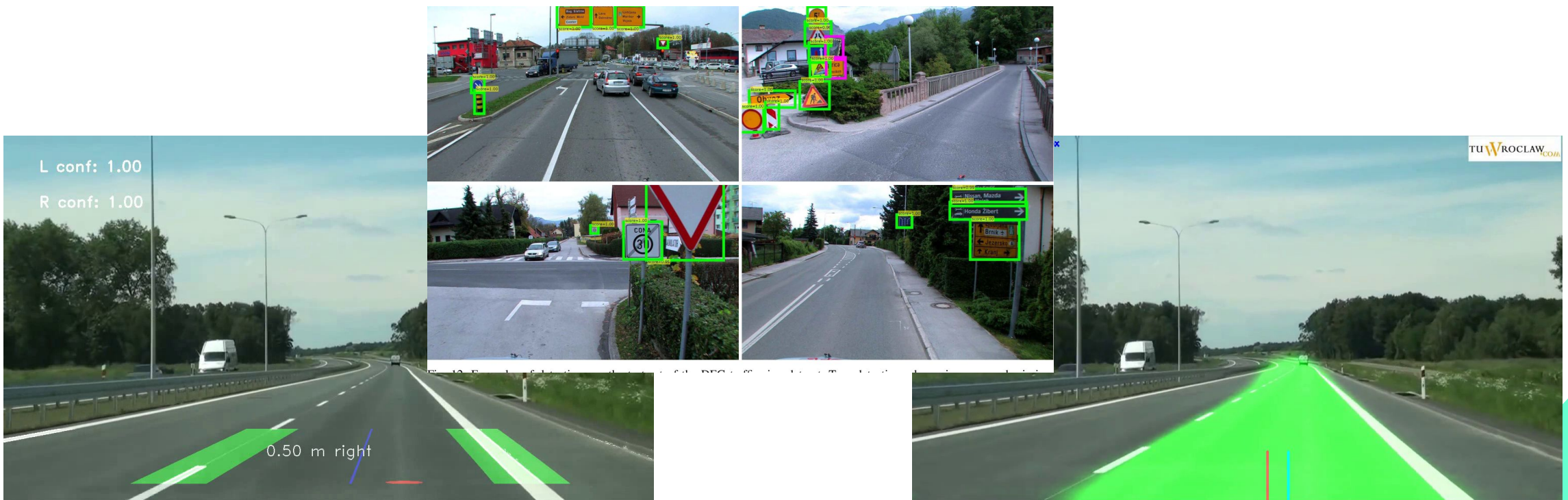
Zadania



1. Wyszukanie i wstępna analiza różnych metod rozpoznawania znaków drogowych i linii.
2. Implementacja jednej lub dwóch wybranych metod (może być metoda "konwencjonalna" i używająca głębokich sieci neuronowych).
3. Weryfikacja działania na filmach np. z kamery samochodowej, dołączenie API z odpowiednimi komendami np. "skręć lekko w lewo", "zwolnij", "ustąp pierwszeństwa za 50 metrów".

Wyniki

Algorytm powinien działać na dowolnym filmie z drogi, załadowanym np. z kamery samochodowej, podobnie jak na obrazkach poniżej:





Pomysły do dalszego rozwoju projektu:

- Detekcja innych pojazdów
- Detekcja pieszych
- Uwzględnianie różnych warunków pogodowych i stanu dróg
- Porównanie różnych metod

Propozycja tematu 3: Smog w Krakowie





Przedstawienie problemu

- Powietrze w Krakowie jest jednym z najbardziej zanieczyszczonych w Europie.
- W Polsce notujemy rocznie około 40 tysięcy nadmiarowych zgonów z powodu jakości powietrza!
- Do badania czynników kryjących się za złym stanem powietrza, należy użyć rozproszonej i licznej sieci czujników.





Główne wyzwania:

- Analiza Time-series
- Duża ilość danych
- Analiza danych geolokalizacyjnych



Inspiracje:

- Czy potrafimy przewidywać jakość powietrza na podstawie innych danych?
 - Jakie czynniki najbardziej wpływają na zanieczyszczenie powietrza?
 - Jak pogoda wpływa na jakość powietrza?
 - Jak bardzo na zanieczyszczenia mają wpływ samochody, fabryki czy elektrociepłownie?
 - Kraków zanieczyszcza okoliczne gminy, czy odwrotnie?
 - W jakich godzinach i gdzie najlepiej uprawiać sport na zewnątrz?
- 
- 



Dane

Zbiór ten zawiera dane o jakości powietrza (stężenia pyłu zawieszonego PM1, PM2,5 i PM10, temperatura, ciśnienie i wilgotność powietrza) z 2017 roku generowane przez sieć 56 sensorów zlokalizowanych w Krakowie.

Pomiary są pogrupowane w 12 plikach, po jednym dla każdego miesiąca.

<https://www.kaggle.com/datascienceairly/air-quality-data-from-extensive-network-of-sensors>



Problemy ze zbiorem danych:

- PM1 nie jest skalibrowany i dlatego może być większy niż PM2,5
- PM2,5 może być większy niż PM10 w granicach błędu pomiaru
- Przez pierwsze dwa miesiące wilgotność i temperatura nie były kalibrowane i dlatego mogą pokazywać niedokładne wartości



Pomysły do dalszego rozwoju projektu:

- Pozyskanie informacji o rodzaju zabudowań
- Pozyskanie informacji o natężeniu ruchu
- Pozyskanie informacji z innych lat, innych czujników

Propozycja tematu 4: Lotnictwo





Przedstawienie problemu:

Każdego roku na świecie odbywa się wiele przelotów, zarówno pasażerskich jak i towarowych. Dodatkowa minuta spędzona w powietrzu oznacza koszty związane z paliwem, czy wynagrodzeniem załogi. Spóźnienia natomiast powodują nieprzyjemności zarówno dla pasażerów, jak i przewoźników.

Czy można je przewidzieć?



Główne wyzwania:

- Analiza dużej ilości danych
- Dane pochodzące z różnych źródeł
- Konieczność wyszukania dodatkowych źródeł danych
- Dane wrażliwe
- Modele grafowe



Inspiracje:

- Czy jesteśmy w stanie przewidzieć czy samolot będzie opóźniony?
- Czy aby na pewno trasy pokonywane przez samolot są optymalne?
(wiadomo linie lotnicze nas oszukują i latają po łuku, zamiast w linii prostej)
- Jakie czynniki wpływają na długość lotu?
- Jak loty oddziałują na siebie nawzajem?
- Czy opóźnienia zależą bardziej od lotniska docelowego, czy startowego?



Dane



Zbiór składa się z trzech plików csv:

- Linie lotnicze
- Lotniska
- Loty

Dane są stosunkowo kompletne

<https://www.kaggle.com/usdot/flight-delays?select=flights.csv>



Dane o lotach



Dane o lotach zawierają:

- **YEAR, MONTH, DAY, DAY_OF_WEEK**: dane dotyczące daty lotu
- **AIRLINE**: identyfikator linii lotniczej
- **ORIGIN_AIRPORT** and **DESTINATION_AIRPORT**: identyfikatory portów lotniczych
- **SCHEDULED_DEPARTURE** and **SCHEDULED_ARRIVAL**: planowany czas odlotu/przylotu
- **DEPARTURE_TIME** and **ARRIVAL_TIME**: prawdziwy czas odlotu/przylotu
- **DEPARTURE_DELAY** and **ARRIVAL_DELAY**: opóźnienie w minutach
- **DISTANCE**: dystans w minutach



Pomysły do dalszego rozwoju projektu:



- Pozyskanie danych pogodowych
- Pozyskanie danych o trasach, jakie pokonują samoloty
- Pozyskanie danych o tym, jakie konfiguracje pasów startowych są dostępne w danym momencie



Plan na resztę dnia pierwszego:



Pozostało około 3 godzin:

1. Dobór zespołów – podział na pokoje na clickmeetingu
2. Wybór tematu – w momencie wyboru dyskusja z trenerem
3. Wybór i wstępna analiza danych – wsparcie w zakresie przetwarzania danych przez trenera
4. Określenie zakresu i działania aplikacji

Plan na sesję samodzielną:

- Dalsza część analizy, stworzenie prezentacji która wyjaśnia jaki temat został wybrany, jakie ma uzasadnienie społeczne, biznesowe, jakie zbiory danych zostaną wykorzystane.
- Może być w formie pitch-decku max 7 minut
(<https://marketingibiznes.pl/start-up-zone/pitch-deck/>)
- Koniecznie nazwijcie swoje projekty i wybierzcie sobie jakieś logo!

Dzień 2





Plan na dzień drugi:

1. Na początku prezentacja grup – jaki temat wybrali, jakie ma uzasadnienie społeczne, biznesowe, jakie zbiory danych zostaną wykorzystane. Prezentacja 7 minut + 3 minuty odpowiadania na pytania reszty grupy.
 - Pytania powinny być maksymalnie “niszczące” należy postarać się znaleźć wszystkie słabe strony danego pomysłu.
2. Szczegółowe konsultacje z trenerem dotyczące dotychczasowych prac i uzgodnienie kolejnych kroków takich jak:
 - analiza danych,
 - wybór modeli,
 - uczenie

Finalnie pod koniec dnia powinny pojawiać się pierwsze wyniki, demo aplikacji.



Dzień 3





Plan na dzień trzeci:

1. Prezentacja postępów przed trenerem w grupach
 - a. jakie modele zostały wybrane,
 - b. jakie pierwsze wyniki,
 - c. co sprawiło problem,
 - d. co można poprawić
2. Dalsza praca nad aplikacją i modelami z pomocą trenera

Sesja samodzielna:

Dalsza praca nad aplikacją i modelami.

Dzień 4





Plan na dzień czwarty:

1. Dalsze prace nad projektami z konsultacjami z trenerem.
2. Dwie godziny przed końcem zajęć pomału zaczynamy pracę nad prezentacjami.
3. Godzinę przed końcem rozpoczynamy prezentację: Nowy dopracowany pitch-deck + prezentacja techniczna. Co zostało zrobione, przy użyciu jakich modeli, jakie są rezultaty. Demo działania aplikacji.



Koniec

