



# Capstone Proposal

Hello,

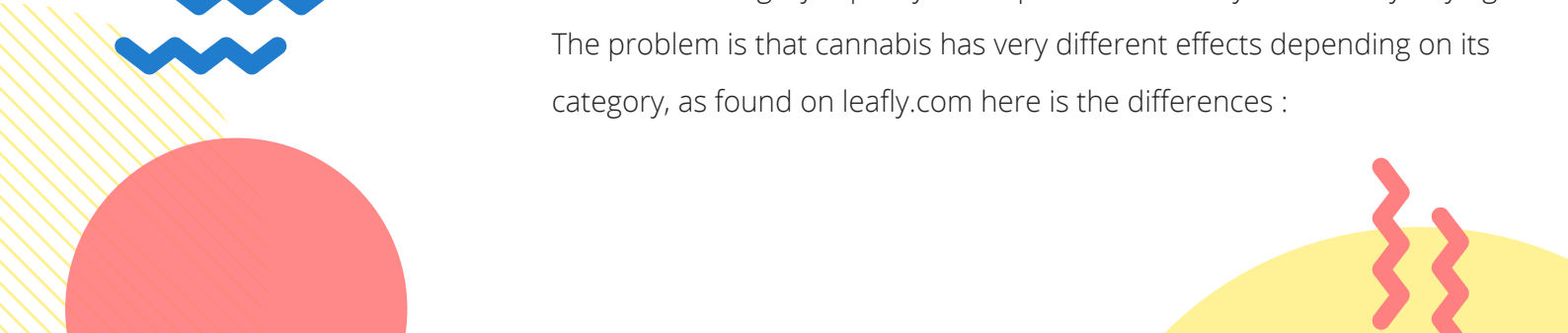
My name is Pierre-Eric, and I'm glad to introduce you my capstone proposal of my project called 420. 420 will be a CNN able to recognize cannabis category (hybrid, sativa, indica) from a simple photo !

## Domain Background

Due to recent law regulations, cannabis become more and more accessible, and the business is booming said Forbes (<https://www.forbes.com/sites/monazhang/2017/11/07/global-marijuana-market-31-billion-investors-cautious/#324df95f7297>). Despite that, cannabis black market remain strong and present. The problem of black market is that you don't always know what's your actually buying. And depending on its category (sativa / indica / hybrid) cannabis can have very different effects. That's why being able to distinguish cannabis category just using a picture could be a real help for people living in countries where cannabis is not legal yet.

## Problem Statement

Even if cannabis going to be legalized in a lot of countries soon, the majority of cannabis consumers are still obliged to access it via black markets. In 2005, cannabis black market was equal to \$141.80 Billion (<https://www.havocscope.com/global-marijuana-market-value-and-users/>). That means that the majority of cannabis consumer don't know what is the category / quality of the product that they are actually buying. The problem is that cannabis has very different effects depending on its category, as found on leafly.com here is the differences :



### PROJECT TYPE

Computer Vision

### DOMAIN

Cannabis recognition

roduce you my capstone  
be a CNN able to recognize  
om a simple photo !

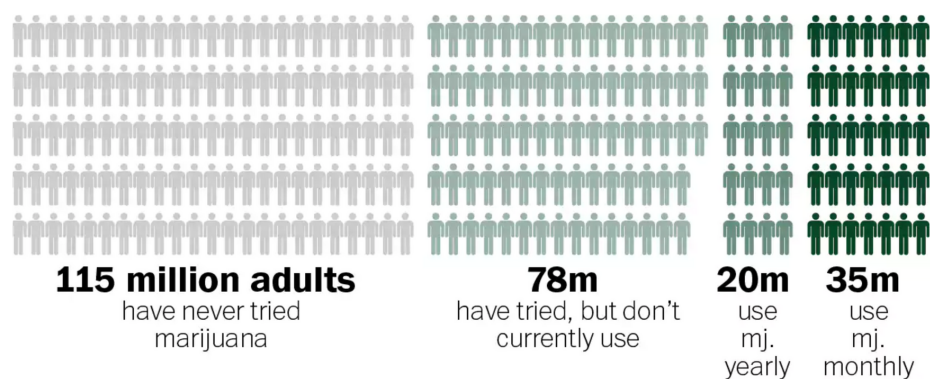
# Capstone Proposal

## PROJECT TYPE Computer Vision

## DOMAIN Cannabis recognition

- Indicas strains are believed to be physically sedating, perfect for relaxing with a movie or as a nightcap before bed.
- Sativas tend to provide more invigorating, uplifting cerebral effects that pair well with physical activity, social gatherings, and creative projects.
- Hybrids are thought to fall somewhere in between the indica-sativa spectrum, depending on the traits they inherit from their parent strains.

That means that nearly 35 millions of american don't know the effect that cannabis will have on them every month (following this chart found on the washington post).



Being able to tell them what effect will have the cannabis they're holding can have great benefits on these people. It will make them more aware about the product they're consuming and will consume it accordingly its effects.



# Capstone Proposal

**PROJECT TYPE**  
Computer Vision

**DOMAIN**  
Cannabis recognition

## Datasets and Inputs


I have collected more than 21,000 images of cannabis on [www.leafly.com](http://www.leafly.com). These images were public and accessible. I have labelled these images by category and strain. Even if I'll not use the strain variable for this project I find it useful to keep it for maybe a 2.0 version. I have collected approximately 8,000 photos of hybrid cannabis, 8,000 photos of indica cannabis and 5,000 photos of sativa cannabis. I will use these images to train a CNN to classify cannabis category.

## Solution Statement

I will build a CNN to classify cannabis category. I will use transfer learning, and try different famous architecture (ResNet50, Inception...) to see what architecture is the best. I'll also apply data augmentation to make my algorithm more adaptative. The final output will be an algorithm being able to classify cannabis category with any cannabis image.

## Benchmark Model

I didn't find any CNN achieving the same goal as mine. Today, the only way to distinguish cannabis category is by looking at it and trying to find specific patterns (<https://herb.co/marijuana/news/indica-vs-sativa-whats-the-difference>).



# Capstone Proposal

## PROJECT TYPE

Computer Vision

## DOMAIN

Cannabis recognition

## Evaluation Metrics

As my classes are not equally distributed, I will use the f-1 score to evaluate my model performance. It considers both the precision  $p$  and the recall  $r$  of the test to compute the score.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

## Project Design

The data collected on [www.leafly.com](http://www.leafly.com) is already labeled and formatted (every picture is on a 100 x 100 pixels format). Here data preprocessing is already done. So the next step will be to use transfer learning and try different architecture. The data passed through the architecture will be augmented in order to make the algorithm more adaptive to future inputs. Once the best architecture for the model has been found I'll try different hyperparameters and different final layer in order to exploit the most of the architecture result.

I hope you've liked my  
capstone proposal,

Pierre-Eric Garcia  
Data Scientist