# Problem

It can be hard to find the right beauty product, so many choices... Although there are a lot of reviews available, people can have different skin types, and usually one won't know whether she/he likes a product until having tried it for a while. If we are buying a product, hoping it can "do something" instead of just looking for a psychological effect, it's probably useful to take a look at its formula. With that, one may ask a few questions:

1. **How important is product formula in determining a beauty product's price?**
   The price of beauty product varies a lot. There is always a debate on whether it is worthwhile to invest on expensive skin cares or makeups. While this could be a matter of personal belief, customers may be curious to know where their money goes for. Is a product expensive because it has sophisticated ingredients, or because it has attractive packaging, or simply because it is from a high-end brand.
   I will try to build a regression model to predict a product's price, and to investigate whether it is the ingredients or other factors that matter the most in determining the price.

2. **Given a product's ingredient list, can we more or less know its category?**
   The cosmetic companies certainly never stop coming up with new concepts. One interesting idea I know is "mask primer" by the brand *Origins*, which is a toner-like product meant to make facial mask more effective. Looking at the ingredients with human eyes, I was not able to tell how this is supposed to be magical.
   Likewise, I have been wondering why people would like to use eye creams: is the skin around eye somehow different from the skin of the rest of the face, and does eye cream has some special ingredients to accommodate that? If not, why would one want to buy an eye cream that is usually of the same price as a face cream but half in size?
   I am curious to see to what extent machine learning can distinguish the category a product belongs to just by looking at its ingredients. Are all different categories really designed for different uses or do they exist simply because of the marketing strategy of cosmetic companies? This can help people decide whether their skin care routine should be:
   *cleaning - toner - serum - moisturizer - eye cream - face oil - mask ....*
   or perhaps only:
   *cleaning - moisturizer,*
   which can help save a lot of time and $$$.

# Data

Data are scraped from beautypedia and paula's choice website.

- Product information from beautypedia (https://www.beautypedia.com), three main tables for skin care, body care and makeup products.
    - name: product name
    - category: subcategory of products
    - brand: product brand
    - ingredient: list of ingredients in a product

- Ingredient information from paula's choice ingredient dictionary (https://www.paulaschoice.com/ingredient-dictionary)
    - name: ingredient name
    - rating: rating of each ingredient according to Paula and her team
    - category: ingredient category -- indicate an ingredient's function in products.
    - description: text description of the ingredient

About Paula Begoun and beautypedia:
Paula Begoun is an American talk radio host, author, and businesswoman. She is known for her view that skin care and cosmetics should be based on ingredients that have been subjected to peer-reviewed research. In 2008, Begoun created Beautypedia.com, where she and her team review beauty products from more than 300 brands.

### Approach

For question 1, we will use regression model. And for question 2, classification would be the best fit. We can also try clustering.

The ingredient list needs data cleaning the most, while the other parts of the data are relatively clean. We will try to match all ingredient to the existing ingredients in the ingredient table. As the ingredient names can have a lot of variations/alias [for example, "alcohol" v.s. "alcohol denat", "Panthenol (Vitamin B-5)" v.s. "Vitamin B5"], we need to design a similarity metric when matching the ingredients.

After the ingredients are matched, we can characterize a formula by various features: how many ingredients in total, how many "good" ingredients, how many antioxidants? ....

## Possible limitations

1. The ratings given by Paula's choice website may or may not be accurate.
2. We only look at individual ingredient. It is possible that certain ingredients working together can have a particular effect, and we won't know this without professional knowledge.
3. Some categories may have too few sample.