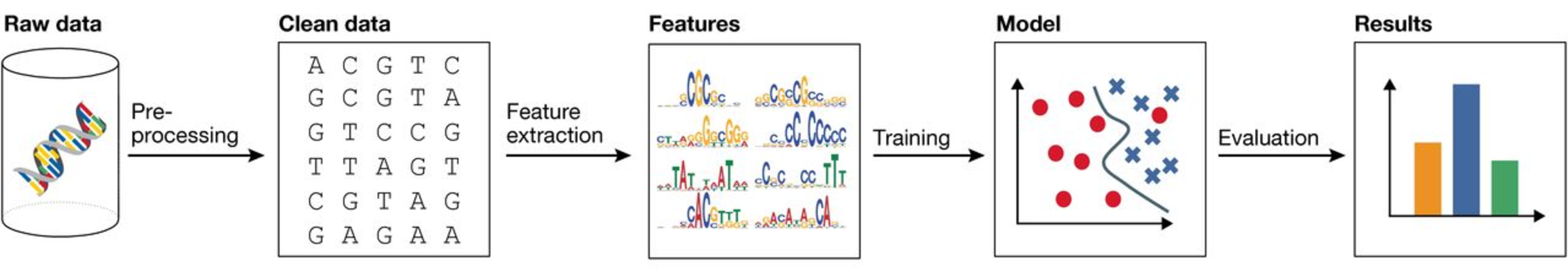
## Mô hình chung cho các bài toán Machine Learning

Phần lớn các bài toán Machine Learning có thể được thể hiện trong hình vẽ dưới đây:

****

Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của tư duy và giao tiếp. (wikipedia)

****

Và bước đầu tiên và không thể thiếu trong việc xử lý ngôn ngữ tự nhiên là tiền xử lý. Vì văn bản vốn dĩ được liệt kê mà không có cấu trúc, để nguyên vậy để xử lý là rất khó khăn. Đặc biệt là loại văn bản trên web có lẫn các HTML tag, code JS, đó chính là noise

### **Tiền xử lý văn bản**

Có rất nhiều cách khác nhau để "làm sạch" và tiền xử lý dữ liệu dạng văn bản. Dưới đây, mình nhấn mạnh một số điểm quan trọng nhất được sử dụng nhiều trong giai đoạn tiền xử lý ngôn ngữ tự nhiên (NLP).

* Xóa thẻ tags: Văn bản chúng ta gặp thường chứa nội dung không cần thiết như các thẻ **HTML**, không có giá trị khi phân tích. Thư viện BeautifulSoup là một công cụ tuyệt vời và cần thiết để xử lý trong trường hợp này.
* Xóa các ký tự có dấu: Trong bất kỳ văn bản nào, đặc biệt nếu bạn đang xử lý ngôn ngữ tiếng Anh, thường các bạn cần phải xử lý các ký tự có dấu. Do đó, chúng ta vần đảm bảo rằng các ký tự này cần được chuyển đổi và chuẩn hóa thành các ký tự ASCII. Một ví dụ đơn giản là chúng ta sẽ chuyển đổi ký tự **é** thành **e**.
* Biến đổi các từ viết tắt: Trong tiếng Anh, các từ viết tắt về cơ bản là phiên bản rút gọn của các từ hoặc âm tiết. Những từ viết tắt của các từ hoặc cụm từ thường được tạo ra bằng cách loại bỏ các chữ cái và âm tiết. Ví dụ như: **do not** -> **don't**, **I would** -> **I'd**. Chuyển đổi từ dạng viết tắt thành dạng đầy đủ cũng là một bước cần thiết để chuẩn hóa văn bản.
* Xóa các ký tự đặc biệt: Các ký tự đặc biệt thường là các ký tự không phải là chữ và số, thường gây "nhiễu" cho dữ liệu của chúng ta. Thông thường, regular expressions **(regexes)** có thể được sử dụng để xử lý vấn đề này.
* Từ gốc và ngữ pháp: Trong các ngữ cảnh khác nhau, các từ gốc thường được gắn thêm các tiền tố và hậu tố vào để đúng với ngữ pháp. Ví dụ các từ: **WATCHES**, **WATCHING**, and **WATCHED**. Chúng ta có thể thấy rằng chúng đều có chung từ gốc là **WATCH**
* Xóa các stopwords: stopwords là các từ có ít hoặc không có ý nghĩa gì đặc biệt khi xây dựng các đặc trưng. Đây thường là những từ giới từ, trợ từ có tần suất xuất hiện tương đối cao trong một văn bản thông thường ví dụ như: a, an, the... Chúng ta không có một danh sách chung tác stopwords tuy nhiên bạn có thể sử dụng thư viện nltk. Hoặc bạn cũng có thể tự xây dựng được các thư viện **stopwords** cho riêng ngôn ngữ bạn đang xử lý. Việc này cũng có ý nghĩa tương đối quan trọng trong xử lý ngôn ngữ tự nhiên.