

Machine Learning Algorithms

Begüm Topçuoğlu
begumtop@umich.edu

Postdoctoral Researcher
Microbiology and Immunology Department
University of Michigan Medical School
Ann Arbor, MI

Outline

- Background
 - What is Machine Learning
- Examples of machine learning algorithms
 - Regression
 - Random Forest
- Code-along ML tutorial.

What is machine learning?

What is machine learning?

Machine Learning: Computer systems learning* input to produce predictions on never-before-seen data using statistical techniques

***Learn:** progressively improve performance on a specific task



“Dog”



“Cat”

Key Terminology

Label: What we are predicting

- IMDB score of a movie
- If a passenger survived Titanic
- If someone has colon cancer

Feature: Input variable

- # of Instagram likes of lead actor
- Age and gender of passenger
- Species abundances in the stool

Maybe just one feature in a simple machine learning project

Many features for a sophisticated machine learning project

Model learns the relationship between the features and the label

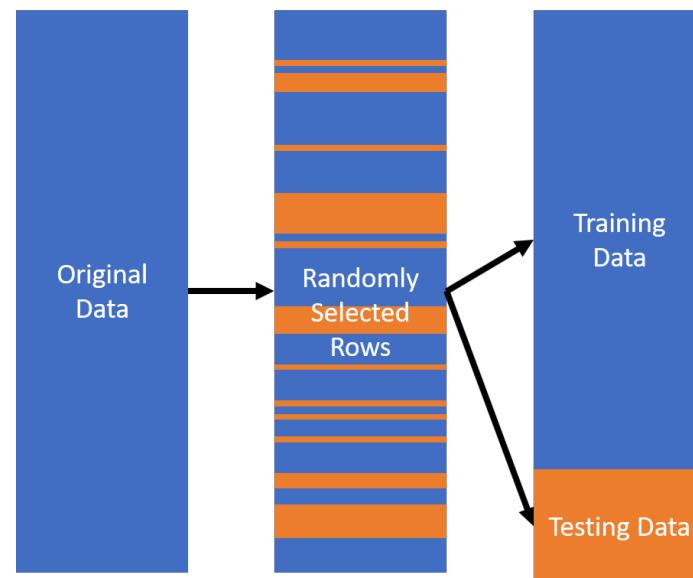
What are other examples where machine learning might be useful?

An example: Is this email spam or not?

- Labels:
 - Spam or Not Spam
- Features:
 - Sender's email address
 - Time of the day it was sent
 - Email contains phrase “Nigeria” or “prince”.
- Model learns the relationship between features and label

Steps in Machine Learning

- Gather Data: Look at hundreds of thousands of emails
- Prepare/clean data: e.g. correct or remove SPAM labeled .edu emails
- Separate data into train and test sets

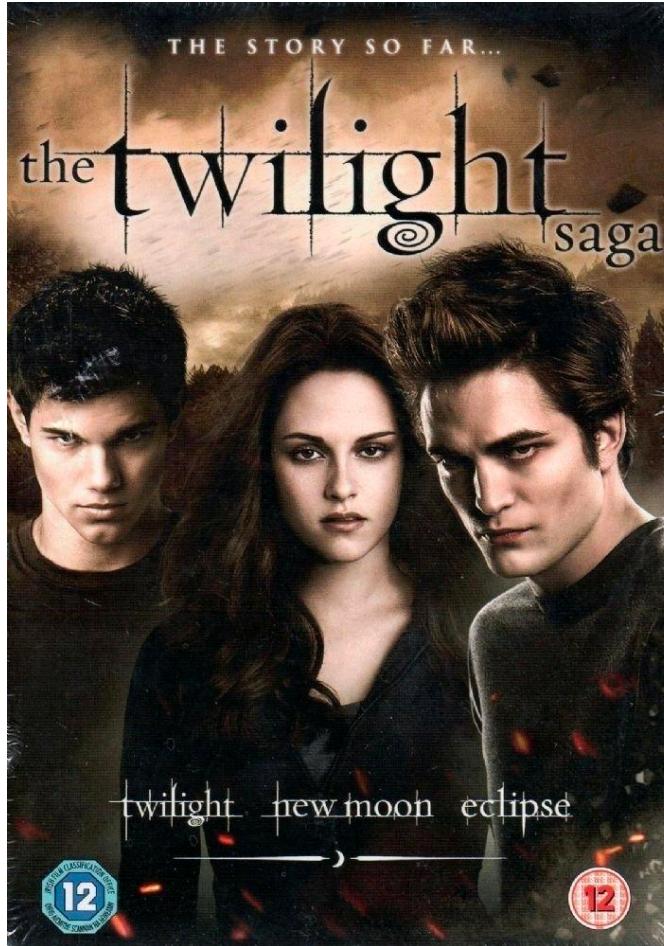


Steps in Machine Learning

- Choose a model: e.g. logistic regression
- Tune parameters: e.g. penalty for getting things too right
- Train (learning): show your model spam email examples, and enable the model to gradually learn the relationships between features and label
- Evaluate: Are we doing well, should we change model/parameters?
- Predict: Apply the trained model to unlabeled emails

Regression

Example: predict IMDB score of a movie

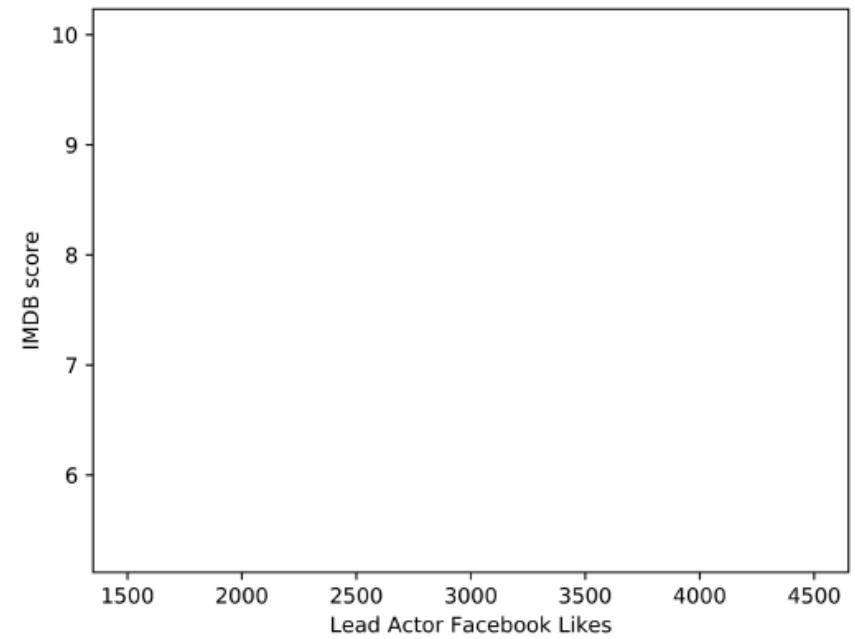


Regression: one feature (simple linear relationship)

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon$$

Label of i^{th} sample
(ex: movie i)

Feature (input)
(ex. lead actor facebook likes)



Regression: one feature (simple linear relationship)

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon$$

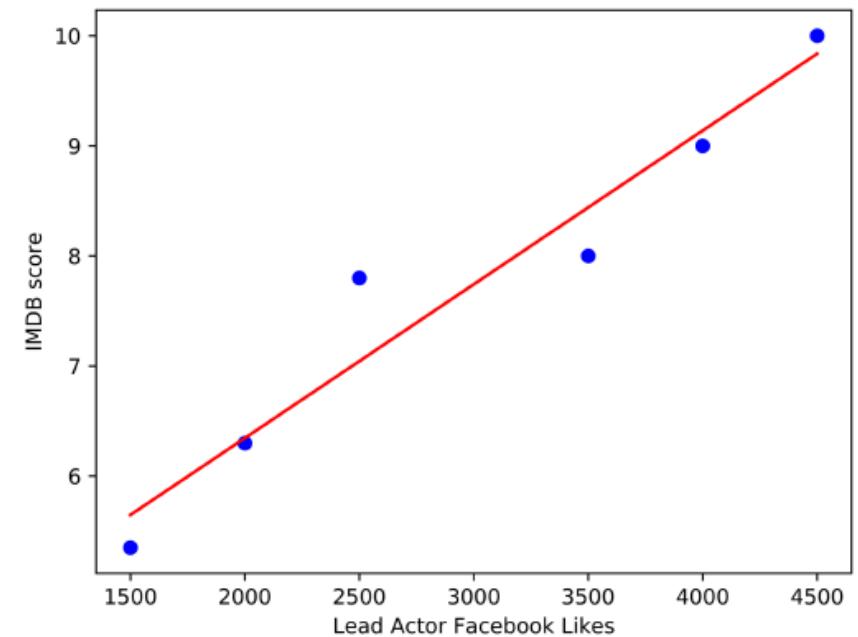
Label of i^{th} sample
(ex: movie i)

intercept

Weight of feature 1

Error

Feature (input)
(ex. lead actor facebook likes)

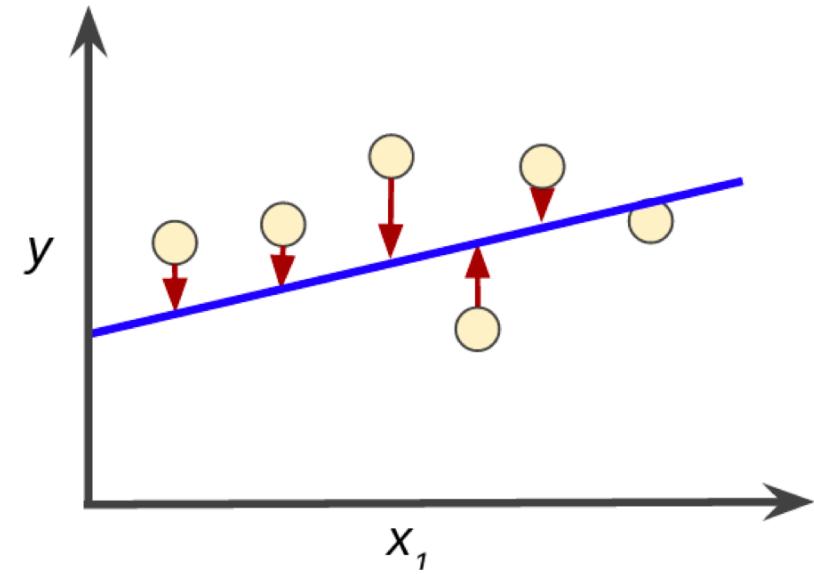
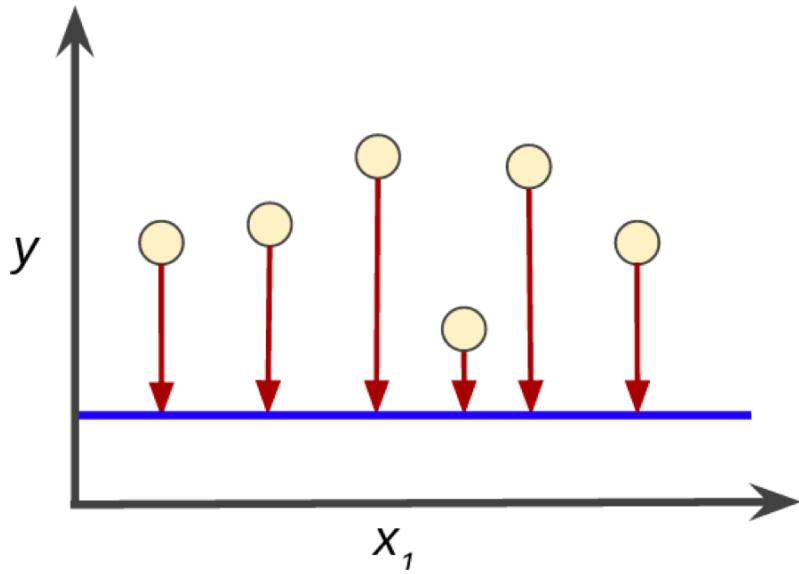


Regression: many feature (more sophisticated model)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} + \epsilon$$

- Features: Country of the movie, genre, duration, plot keywords, budget etc.
- Each of these features will have different weights.
- Training a model simply means learning (determining) good values for all the weights from labeled examples.
- How does a model do that?

The model learns by making bad predictions and getting a penalty for the bad prediction



- The goal of training a model is to find a set of weights that have low loss, on average, across all examples (minimize the total error).

Be careful of overfitting!



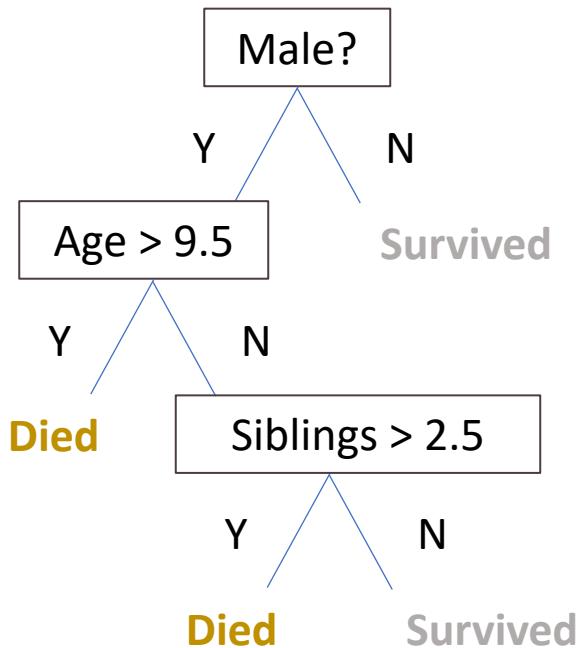
Underfitti

Overfitting

Random Forest

Titanic Data Set

Decision Tree

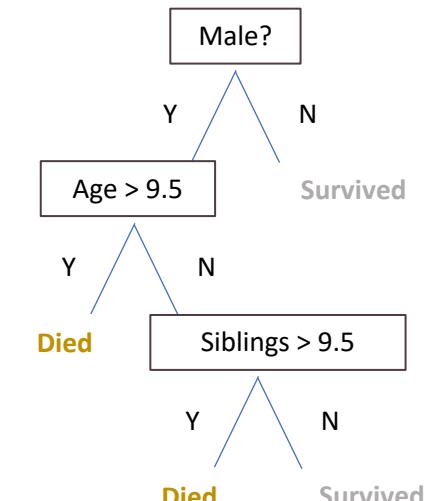


Overfitting

Random Forest

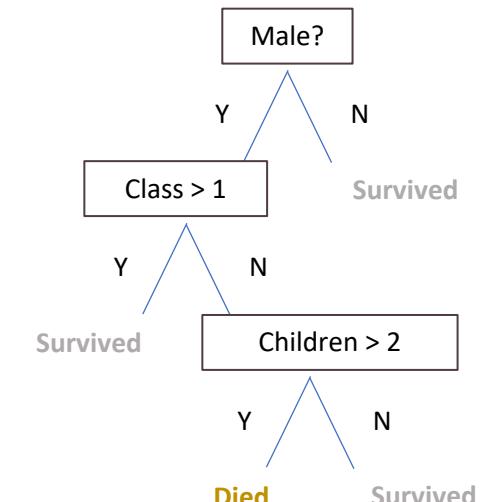
Random Forest

Name	Class	Sex	Age	Siblings/ Spouses Aboard	Parents/ Children Aboard	Fare	Survived
Mr. Owen Harris Braund	3	male	22	1	0	7.25	0
Mrs. John Bradley (Florence Briggs Thayer) Cumings	1	female	38	1	0	71.2833	1
Miss. Laina Heikkinen	3	female	26	0	0	7.925	1
Mrs. Jacques Heath (Lily May Peel) Futrelle	1	female	35	1	0	53.1	1
Mr. William Henry Allen	3	male	35	0	0	8.05	0
Mr. James Moran	3	male	27	0	0	8.4583	0
Mr. Timothy J McCarthy	1	male	54	0	0	51.8625	0
Master. Gosta Leonard Palsson	3	male	2	3	1	21.075	0
Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	3	female	27	0	2	11.1333	1
Mrs. Nicholas (Adele Achem) Nasser	2	female	14	1	0	30.0708	1
Miss. Marguerite Rut Sandstrom	3	female	4	1	1	16.7	1
Miss. Elizabeth Bonnell	1	female	58	0	0	26.55	1
Mr. William Henry Saundercock	3	male	20	0	0	8.05	0
Mr. Anders Johan Andersson	3	male	39	1	5	31.275	0
Miss. Hulda Amanda Adolfina Vestrom	3	female	14	0	0	7.8542	0
Mrs. (Mary D Kingcome) Hewlett	2	female	55	0	0	16	1
...



Random Forest

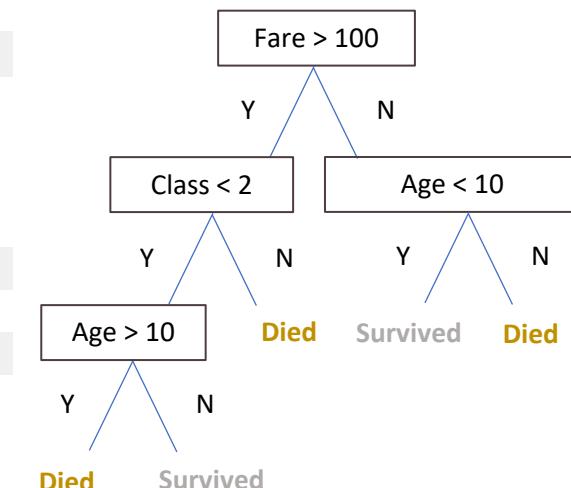
Name	Class	Sex	Age	Siblings/ Spouses Aboard	Parents/ Children Aboard	Fare	Survived
Mr. Owen Harris Braund	3	male	22	1	0	7.25	0
Mrs. John Bradley (Florence Briggs Thayer) Cumings	1	female	38	1	0	71.2833	1
Miss. Laina Heikkinen	3	female	26	0	0	7.925	1
Mrs. Jacques Heath (Lily May Peel) Futrelle	1	female	35	1	0	53.1	1
Mr. William Henry Allen	3	male	35	0	0	8.05	0
Mr. James Moran	3	male	27	0	0	8.4583	0
Mr. Timothy J McCarthy	1	male	54	0	0	51.8625	0
Master. Gosta Leonard Palsson	3	male	2	3	1	21.075	0
Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	3	female	27	0	2	11.1333	1
Mrs. Nicholas (Adele Achem) Nasser	2	female	14	1	0	30.0708	1
Miss. Marguerite Rut Sandstrom	3	female	4	1	1	16.7	1
Miss. Elizabeth Bonnell	1	female	58	0	0	26.55	1
Mr. William Henry Saundercock	3	male	20	0	0	8.05	0
Mr. Anders Johan Andersson	3	male	39	1	5	31.275	0
Miss. Hulda Amanda Adolfina Vestrom	3	female	14	0	0	7.8542	0
Mrs. (Mary D Kingcome) Hewlett	2	female	55	0	0	16	1
...



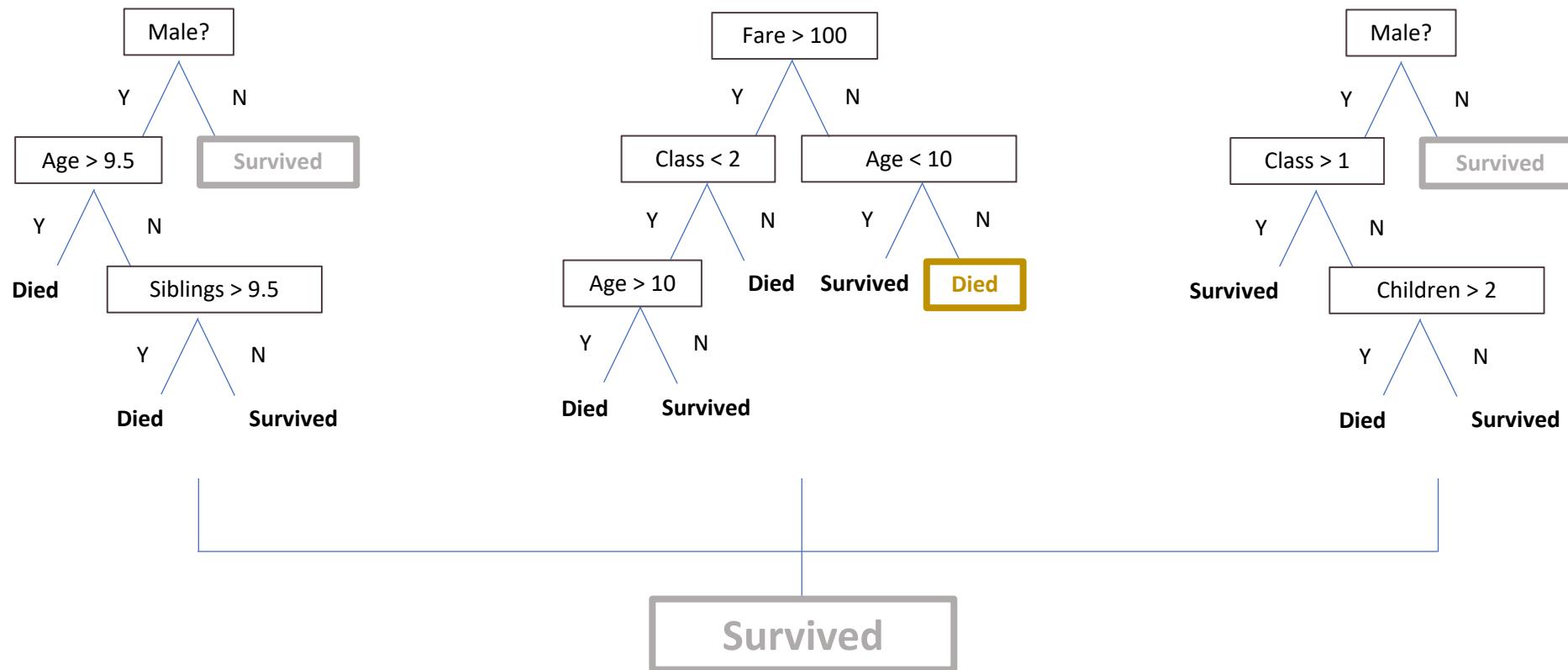
Random Forest

Name	Class	Sex	Age	Siblings/ Spouses Aboard	Parents/ Children Aboard	Fare	Survived
Mr. Owen Harris Braund	3	male	22	1	0	7.25	0
Mrs. John Bradley (Florence Briggs Thayer) Cumings	1	female	38	1	0	71.2833	1
Miss. Laina Heikkinen	3	female	26	0	0	7.925	1
Mrs. Jacques Heath (Lily May Peel) Futrelle	1	female	35	1	0	53.1	1
Mr. William Henry Allen	3	male	35	0	0	8.05	0
Mr. James Moran	3	male	27	0	0	8.4583	0
Mr. Timothy J McCarthy	1	male	54	0	0	51.8625	0
Master. Gosta Leonard Palsson	3	male	2	3	1	21.075	0
Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	3	female	27	0	2	11.1333	1
Mrs. Nicholas (Adele Achem) Nasser	2	female	14	1	0	30.0708	1
Miss. Marguerite Rut Sandstrom	3	female	4	1	1	16.7	1
Miss. Elizabeth Bonnell	1	female	58	0	0	26.55	1
Mr. William Henry Saundercock	3	male	20	0	0	8.05	0
Mr. Anders Johan Andersson	3	male	39	1	5	31.275	0
Miss. Hulda Amanda Adolfina Vestrom	3	female	14	0	0	7.8542	0
Mrs. (Mary D Kingcome) Hewlett	2	female	55	0	0	16	1
...

Survived



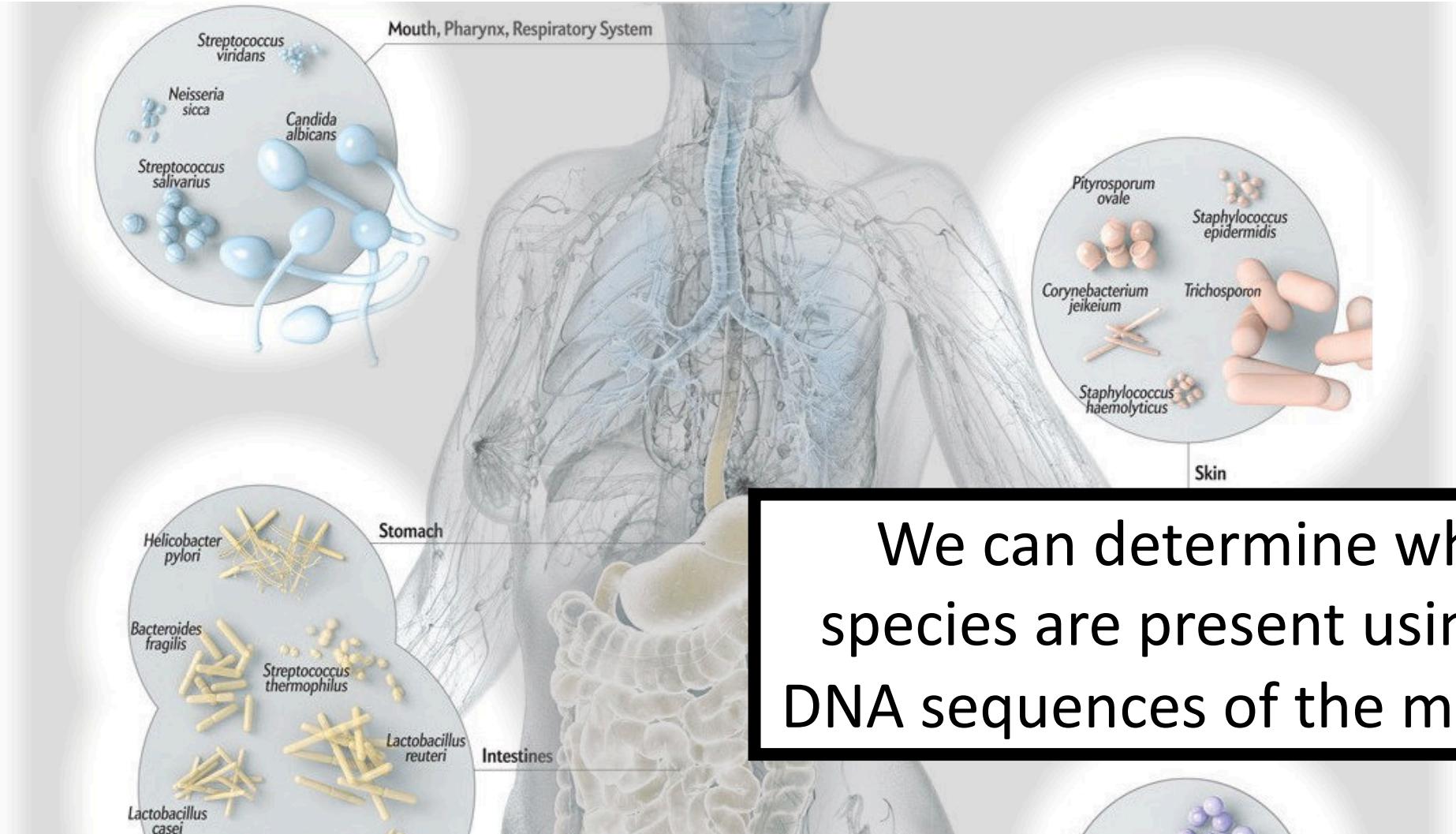
Name	Class	Sex	Age	Siblings/ Spouses Aboard	Parents/ Children Aboard	Fare	Survived
Miss. Ellen O'Dwyer	3	female	24	0	0	7.8792	1



Example: Predict if lesion is cancerous
(we'll be doing this in the lab!)

- Labels:
 - Colorectal lesions of patients
 - Defined as cancer or not
- Features:
 - 16S rRNA gene sequences in stool

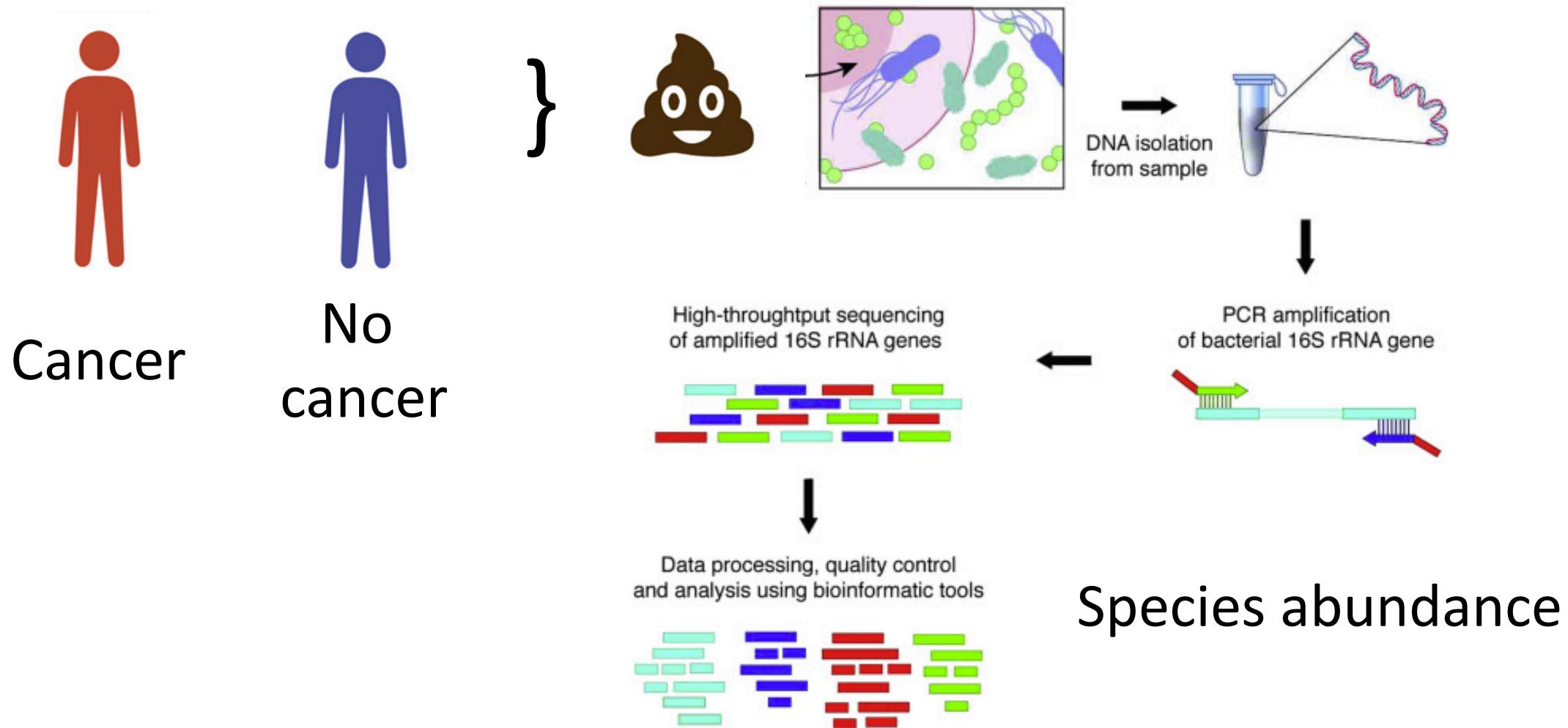
We all have microbes in us: *the human microbiome*



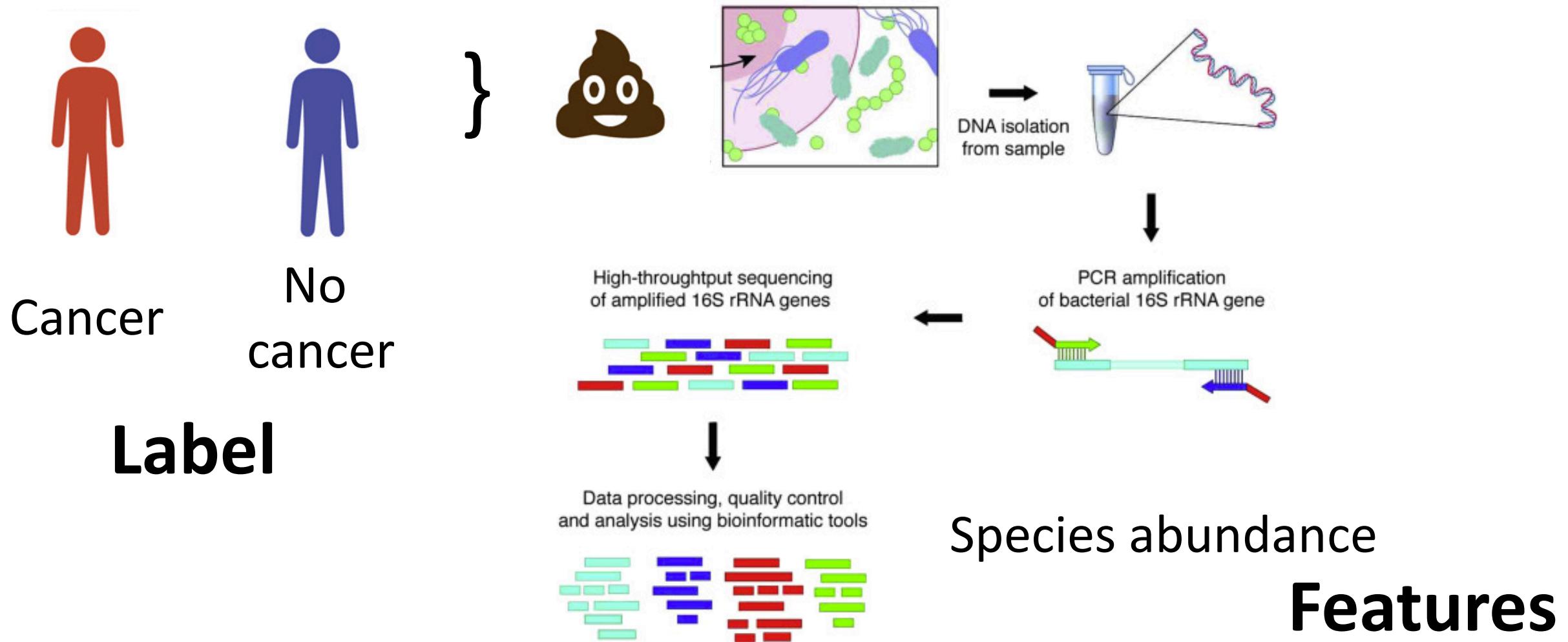
We can determine which species are present using the DNA sequences of the microbes

Can you use the microbes in you to predict if you have/will get colorectal cancer?

Can you use the microbes in you to predict if you have/will get colorectal cancer?



Can you use the microbes in you to predict if you have/will get colorectal cancer?



Download data

- Download this data:
 - <https://tinyurl.com/yyqywozj>
- Make a folder on your Desktop called mtbi
- Make a folder in mtbi called data
- Move data.tsv to the data folder