# Capstone Project

Nuestro Universo está poblado por miles de millones de galaxias, cada una formada por miles de millones de estrellas. Los telescopios, situados tanto en satélites como en tierra, observan el cielo y toman una enorme cantidad de fotos digitales que luego los científicos estudian para entender cómo nacen las galaxias, cómo interaccionan entre ellas, y cómo el Universo en su conjunto se ha formado, cómo va cambiando y así descubrir las leyes que dictan su evolución.

Hemos seleccionado para este “Capstone Project” un conjunto de datos perteneciente a un proyecto llamado “Galaxy Zoo”. A lo largo de este último curso os guiaremos para que podáis analizarlo con una parte de las herramientas Big Data que habéis aprendido en cursos anteriores, y que nos ayudará a explorar y conocer el Universo en el que vivimos.

## Las galaxias

El Sol es una de las muchas estrellas que forman parte de nuestra galaxia: la Vía Láctea. Observando las estrellas más lejanas podemos ver cómo se agrupan formando galaxias, como la Vía Láctea, de distintas forma y tamaño. Su forma, tamaño y brillo nos revelan cómo se forman y evolucionan.

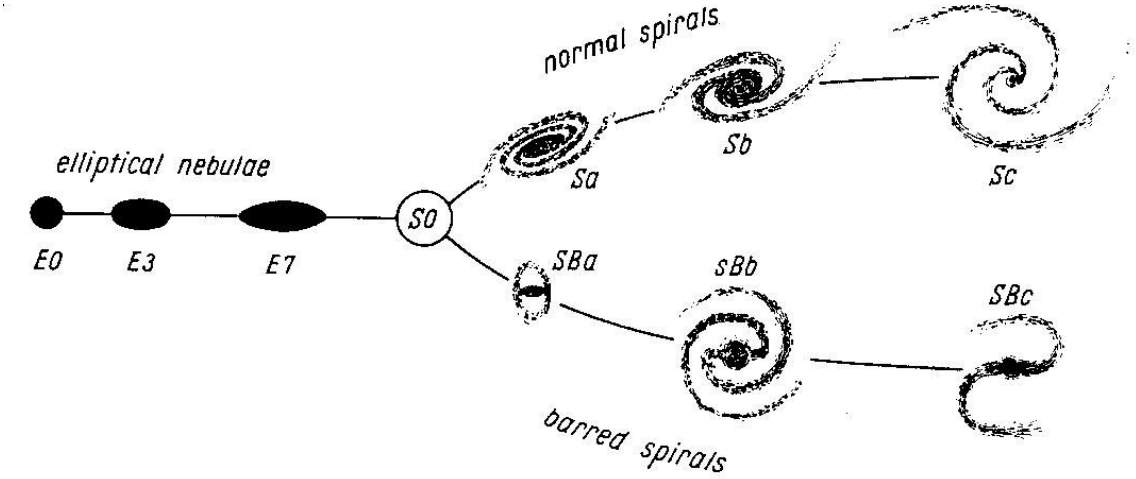
## Astronomía y Big Data

Potentes telescopios escanean el cielo y recogen fotos en formato digital de estrellas y galaxias lejanas. Son capaces de recoger imágenes de objetos celestes que a simple vista no podríamos observar. Aun así solo tenemos información de una pequeña fracción de la enorme cantidad (centenares de miles de millones) de galaxias que pueblan nuestro Universo. La Cosmología es la ciencia que estudia cómo el Universo ha nacido y evolucionado y así poder entender cuál podría ser su destino. Conocer los distintos tipos de galaxias y clasificar su forma es uno de los pasos fundamentales para el avance en el conocimiento del Universo en que vivimos.

**La clasificación de imágenes de galaxias basada en su forma es el objetivo del proyecto final de esta especialización**. Para alcanzarlo tendréis la oportunidad de aplicar algunas de la técnicas de análisis y clasificación de Big Data que habéis ido conociendo durante los cursos y las semanas pasadas.

# Clasificación

Existen muchas maneras de clasificar las galaxias (por ejemplo por el color, o el brillo) pero la más común es por su foma, que depende de su edad, composición, etc., según el esquema de Hubble que podéis ver aquí abajo.



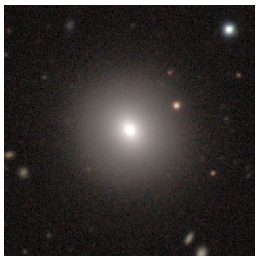

Clasificar una galaxia por su forma no es siempre tarea fácil. Los tipos más reconocibles de galaxias son las espirales y las elípticas, pero hay muchísimos estados intermedios, o algunas con forma irregular que, junto con el tamaño, la orientación del objeto y la resolución de la imagen, pueden dificultar el trabajo del clasificador. Al contrario que el color, el brillo o la distancia, que son propiedades cuantitativas y pueden medirse de manera directa, la forma es mucho más difícil de valorar y depende de criterios más subjetivos.

# GalaxyZoo

El proyecto GalaxyZoo consiste en recolectar datos sobre la forma del mayor número posible de objetos celestes. Para llevar a cabo esa tarea el proyecto cuenta con la colaboración de voluntarios que, a través de una página web, visualizan imágenes en su ordenador personal y clasifican el objeto fotografiado. El resultado se envía a través de la misma página web.

En la página de GalaxyZoo https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/ se puede acceder a un tutorial, en inglés. Los objetos representados en las imágenes se clasifican según los siguientes criterios:

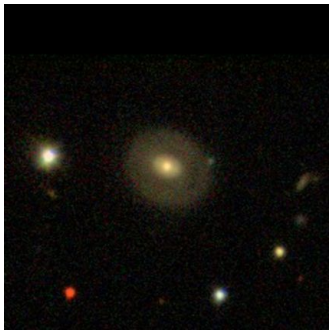
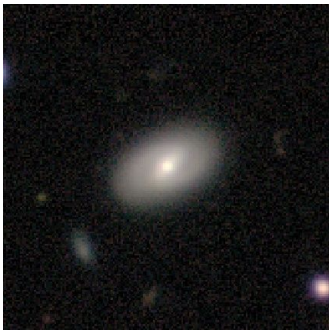
a. Una galaxia ELÍPTICA, cuyo brillo va disminuyendo gradualmente desde el centro de la imagen



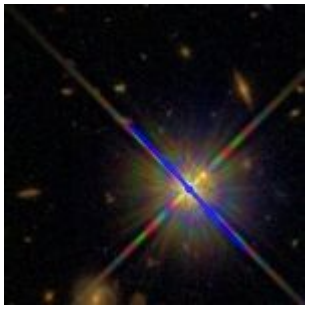
b. La imagen tiene estructuras, que pueden ser:

- los brazos de una galaxia ESPIRAL:

- un núcleo, o unas barras, u otras características peculiares:

- una estrella, o una traza de satélite, o algún otro artefacto que obstaculice la posibilidad de clasificar el objeto:




En el proyecto GalaxyZoo original, voluntarios tenían que distinguir entre objetos elípticos, espiral horaria, espiral antihoraria, casos intermedios, estrella/no identificable, o fusión de

objetos. Para este proyecto vamos a adoptar una versión de clasificación más simple, y únicamente distinguiremos entre objetos de forma elíptica, o dicho de otra forma, sin estructuras, y objetos con forma espiral, caracterizados por la presencia de estructuras.

## Datos

Los datos que se van a utilizar en este proyecto son:

- un subconjunto de imágenes de galaxias tomadas por el telescopio de un proyecto llamado Sloan Digital Sky Survey (SDSS)
- un sub-conjunto de parámetros asociados a cada galaxia (identificador único en el catálogo de SDSS, posición en el cielo, brillo, distancia, etc.)
- un subconjunto de los resultados de la clasificación web hecha a través del proyecto GalaxyZoo.

Con estos datos y con la herramientas que ya han sido presentadas en el curso de esta especialización, te guiaremos en las próximas semanas para que puedas desarrollar y finalmente presentar a un imaginario comité científico tu método de clasificación y análisis de galaxias basados en herramientas Big Data para poderlo aplicar a las miles de millones de galaxias observadas.

## Herramientas

La herramientas que utilizaremos son:

- HDFS y sus comandos de consola para la ingestión de datos.
- Hive y su cliente beeline para la creación del modelo de datos, la importación de los datos externos, su exploración preliminar y su análisis posterior.
- Spark y python notebooks para el análisis, visualización e interpretación de los resultados.

Estas herramientas se van a proporcionar mediante la máquina virtual de Cloudera que ya habéis utilizado en los cursos anteriores.

En resumen: os guiaremos para que crees un clasificador de imágenes, para que se pueda aplicar a un volumen grande datos.  En los ejercicios propuestos en la sección “Trabajo a realizar” tendréis que trabajar con la máquina virtual y las herramientas que habéis ido aprendiendo a lo largo de los cursos anteriores. Esos ejercicios son preparatorios para poder contestar a las preguntas del quiz evaluable.

En la última semana os guiaremos para preparar el resumen final del trabajo hecho en este capstone project. Será evaluado por otros estudiantes de este mismo curso (peer review), y además tendréis que evaluar vosotros el trabajo de otros compañeros.