Domain

Fictional characters, especially cartoon characters, often have characteristic ways of speaking. In this project I will attempt to train an algorithm to predict the speaker of an unknown piece of dialogue by training it on tagged examples.

In the MVP version, I'll focus on the main cast (5-8 characters) of a single show, Futurama. I've chosen Futurama because I'm (very) familiar with it, and I know that its characters speak in characteristic ways (e.g. Farnsworth is rambling, bitter, and uses lots of science(y) words; Kif is servile and dots on his smizmar Amy; Bender is sociopathic and likes to talk about theft and alcohol).

Data

The minimal dataset will be everything that's been said by these characters over the course of 10 seasons and a hand-full of movies. Time permitting, I'll train the algorithm on character dialogue from other works to compare separability.

With the exception of a few episodes with oddly formatted transcripts, I managed to gather most of the this data on Wednesday by scraping TheInfoSphere.org. This comes to 24,034 lines of dialogue and is in the format of the below `DataFrame`.

```
In [35]:
1 df = pd.DataFrame(data)
2 df.sample(5)
```

Out[35]:

	ep_code	ep_title	location	speaker	text
1862	S01E09	Hell Is Other Robots	11	Beastie Boys	[singing] Well it's 50 cups of coffee and you know it's on\nI move the crowd to the break of bre...
21319	S08E03	Ghost in the Machines	183	Bender's ghost	Shut up, God!
13518	S05E08	The Why of Fry	29	Bender	Ah, buck up, meatloaf. Bender'll take you out tonight and cheer you up. What do you wanna do? An...
7278	S03E05	Amazon Women in the Mood	78	Kif	Is there nothing we can do, sir?
8393	S03E06	Bendless Love	173	Bender	Admit it. You felt something for me tonight. And by "me" I mean Flexo.

Known unknowns

- How separable the dialogue is theoretically, given a perfect model.
- Whether this is enough dialogue to train a model.
- The number of words needed to make a piece of dialogue characteristic.