Final Project

You have two options for your final project; please pick either (or both)!

Option A - Audio Extrapolation

One of the labs you worked on used multi-output, linear regression to extrapolate the missing part of an audio file. The resulting clip was probably a lot more fascinating than the very poor, linear regression accuracy scores. That low accuracy has actually been bugging us, so we're interested in what you might do to get better sounding audio extrapolated, as well as a higher scoring result.

Hints

- Re-read through the course and try to apply the different you've picked up on the way to increase your accuracy.
- Browse through SciKit-Learn's documentation for algorithms, similar to multi-output linear regression, that might give you better results. Don't limit yourself to what we've covered in the course if you chance upon something that seems better. Nothing is withheld from you. You are free to use any technique you can come up with, including manufacturing your own features, using any machine learning algorithms you can find, or even coding up your own algorithms from scratch.
- Review the assignment and look how large your testing set is—is it big enough? If you need more data, download audacity, learn how to use it, and then record yourself! Feel free to share GitHub links to your own audio 'digits' libraries with each other.
- If you haven't already done this, you should try recording yourself saying '*zero*' in a half-second audio

Wrap-up

Post-course Survey
Survey

clip. Replace one of the training audio files from your lab assignment with your recording, and then hardcode the random_idx value to that clip. That way you'll be able to hear your algorithm attempt to complete your own voice's missing audio!

- Discuss your ideas on the Forum... but don't post any solutions! If you're dying to post something, post a link to your extrapolated audio output for others to hear, or share your digits recording library.

Option B - Astrophysics

Another interesting area we didn't get the opportunity to dive into as deeply as we'd have liked is the realm of astrophysics. There is a machine learning library called astroML, which is built on NumPy, SciPy, MatPlotLib, and Scikit-Learn... all the libraries you're already familiar with from this course. It also uses AstroPy, a library that has much functionality aimed at professional astronomers and astrophysicists, but useful to anyone developing astronomy software.

Astropy offers you the ability to do celestial coordinate and time transformations, allows you to interact with HDF5 files (here's an amazing tutorial in R, but H5Py is Python module you can use to follow along), and much more. In fact, it's used professionally by the High Energy Astrophysics Science Archive Research Center, the National Virtual Observatory, and elsewhere.

On the 12th of November 2014, the European Space Agency, partnering with NASA and others, were able to land a small lander named Philae on a comet. In addition to the lander, another satellite, Rosetta, continues to orbit the comet till today. Since then, a lot of the scientific data has come back to us from Rosetta and Philae comet experiments, and has also been made publicly available for analysis.

Of particular interest is the infrared spectra data from the comet. Scientist use spectroscopy to study the make up of matter through its interaction with light. Wouldn't it be awesome to confirm the research top astrophysicists have done in discerning the composition of the comet 67P/Churyumov-Gerasimenko, using machine learning and data analysis? We think you're up for the challenge!

Hints

- If you aren't familiar with 67P and the Rosetta/Philae pair, start by reading up as much as you can about the both of them. The history of the crafts, what went right, what went wrong, etc. It's always a good idea to be as acquainted with the subject of your data as possible, in lieu of domain expertise and experience.
- Browse the sites linked above and try to procure scientific data from 67P, but don't limit yourself to them. Browse the internet and look around. In this case, you aren't the one who's put together the scientific experiments, but you do have to hunt for the data. Of particular interest is the data originating from the VIRTIS (Visible and Infrared Thermal Imaging Spectrometer) instrument, and the OSIRIS (Optical, Spectroscopic and Infrared Remote Imaging System) instrument.
- The data coming from the comet is really *unknown* data. Before doing anything with it, you need to find a few spectroscopy datasets of known elements here on earth, or of known organic compounds (depending on if you're interested in identifying elements or compounds). Train and test against them until you have a high level of accuracy.
- Run your algorithm against the comet datasets and see what results you come up with!
- This is a challenging problem, particularly finding good data sources. Feel free to share resources on the discussion forum, or to even form teams.

Notes

Please note, this is an ungraded final project! If you're able to complete either them, then you've surely done a fine job of absorbing the material of this course, and have even gone far beyond it. Regardless of your score in the class, you absolutely deserve **full marks** =). We won't actually be able to *award* you full marks, due to the complexity of manually grading tens of thousands of final projects, but you'll know in your heart and so will we.

	Even if you aren't able to complete either version of the final project, it will still be a very difficult and rewarding learning experience if you attempt to do so, and that will only help you towards your goal of becoming a full-fledged data scientist.