My name is Vernon Gayle. I am a professor of Sociology in Social Statistics at the University of Edinburgh. And I am part of the National Centre for Research Methods (NCRM). I hate appearing on video, however in this short presentation I'm going to tell you about Jupyter notebooks and how they can help benefit Social Science research. If you're a social science researcher and interested in improving how you undertake statistical data analysis, in this video, it is designed to introduce you to Jupyter notebooks. I'm a fairly recent convert to use in Jupyter notebooks but in the next 10 minutes I hope to convey some of my enthusiasm and to encourage you to consider using them in your research.

But first I'll start by talking about the social science workflow. Having planned and organized workflow is essential for high-quality statistical research using large scale social surveys or administered in social science data set. The workflow referred to a coordinated framework for conducting social science data analysis. Workflow includes planning, organizing executed and documenting analysis.

J Scott Long has provided an extensive and almost the rabbinical account of good workflow practices. I suggest that any data analyst who has not read Long's book would benefit from doing so whatever their age or their career stage. Central to a successful workflow is the audit trail. Your audit trail is nothing more than a chronological account of the activities undertaken in the data analytical process. Alternatively you could think of the audit trailers as the line of breadcrumbs helping you navigate through the research process. The audit trail is important because within the statistical analysis of social science data set, a minor decision such as dropping some cases from an analysis or re-coding a variable can have major consequences later on. Keeping track of even the most seemingly inconsequential action within the workflow is important i.e. it facilitates transparency and makes contributions to efficiency and accuracy and ultimately to the overall success of the research project.

My acquaintance with Phillips Stark at UC Berkeley says that not having a plan and organize workflow can be compared to drinking and driving. In both cases it doesn't matter how careful you are, it's still highly likely to end in a wreck. In just the same way as I would never advocate drinking and driving, I also don't advocate undertaking a search without a planned and organized workflow. There is a long history in the Natural Sciences of researchers continuously making notes that contribute to high quality documentation. Nobel Prize winner Linus Pauling used these bound notebooks to keep track of the details of his research and the 46 notebook spanning a period of 1922 until 1994 are available online.

Professor Pauling's notebooks include calculations, experimental data, scientific conclusions, ideas for further research and numerous autobiographical reflections. For example notebook 24 on page 151 contains an entry detailing his golden wedding anniversary. But the use of notebooks goes back much further. For example Galileo used notebooks which directly integrated his data for example drawings of Jupiter and its moons with key metadata for example the timings of each observation, the weather and even telescope properties. The data and metadata were annotated and the text which included descriptions of methods, analysis and scientific conclusions were woven together. This links us neatly to Jupiter notebooks. The three Galilean moons are visible in the Jupyter logo.

Why is it called Jupyter? Well the computer languages Julia Python and R almost spelled JuPyter that's why it's called Jupiter. Jupyter notebooks are currently used in big science. The recent detection of the gravitational waves is that is it hailed as a major scientific discovery. If you follow this youtube link you can watch a short video of Fernando Paris who first conceived Jupyter notebooks demonstrated a Jupyter notebook that includes data and analysis of the first gravitational waves detected by the LIGO team.

What are Jupyter notebooks? They are an open-source web application that facilitates the creation and sharing of documents that contain LIVE code and supporting commentary in the form of an explanatory text. It's a platform that can be used throughout the research process to organize an articulate elements of the social science workflow. The Jupyter notebook is open source and supports interactive data analysis in over 40 programming languages.

What do Jupyter notebooks offers social science data analysts? First they facilitate easy documentation alongside research code. They have good portability because notebooks are ready to share. They're language agnostic e.g. an analysis can be undertaken using many different languages. They can produce rich visual outputs, they can leverage big data research tools for example using Python. They can be used as integrated tools in teaching, training, knowledge exchange in research capacity building and finally they support and facilitate collaborative work.

Now let me quickly show you around the Jupyter notebook. Cells can contain three things live research code for example stata or R syntax that can be executed in this case it's a static command. Cell can contain the results of data analysis here we see the output for the static command. And finally cell can contain text comments that form the documentation of the research workflow. Documentation in Jupyter notebooks is relatively easy you can use markdown which is a simple and easy to learn form of plaintext.

Jupyter notebooks are language agnostic. It's possible to work in many languages within a notebook. I'm going to show you a few screen grabs now of a statistical model estimated within a Jupyter notebook.

Within the notebook I've estimated a logistic regression model first in Stata then the same model in R and finally the same model again in Python. These models have all been estimated within the same notebook and this hopefully illustrates that i can simply and easily move between different data analysis programs within a single Jupyter notebook. The next example involves rich text output. I had a great wee group of PhD students a few years ago who absolutely loved XKCD web comics so just for them and they know who they are I've produced the plot in the style of an XKCD graphic to show the graphing ability of Jupyter notebooks. Continuing on the theme of which visual outputs, here is an example that uses an open-source street map. I recently moved to more commodious office around the corner in some place in Edinburgh and here's an example of embedding a map within a Jupyter notebook. The inclusion of maps offers a great deal of potential especially when working with Geo-coded data.

One of the many exciting features of Jupyter notebooks is their interactive facilities. This next example uses image processing to identify galaxies in an image of the sky provided by the Hubble Space Telescope. This is a live example hosted by the journal Nature and after running the cell we can explore the parameters of the detection algorithm to find galaxies of different sizes and prominences.

Just before i finish I want to direct you towards Professor Lorena Barb's website. Lorena demonstrates the value of using Jupyter notebooks in researching from teaching by weaving executable code within multimedia context.

In a nutshell when you use a dupe the notebook for Social Research you end up with an uber Stata .do file or an uber R script that contains your research code and your output woven into a literate research narrative. Jupyter notebooks can be converted into other formats for example PDF or HTML ready for presentation for publication for collaboration and for sharing.

Hopefully this will help convince you of the benefits of using Jupyter notebooks and how they have an obvious appeal to undertaking a reproducible research. It's very easy to install Jupyter and you can get started relatively quickly. Very soon you'll be able to use Jupyter notebook standing on your head well almost. But on a more serious note I'll conclude by saying overall Jupiter notice offer a useful and usable environment in which to plan, organize and execute your data analysis while simultaneously been able to record document and archive your search results alongside the code that produce them. Jupyter notebooks have the capacity to transform how statistical analysis using large scale social surveys or administrative social science data set is routinely undertaken. Good luck.