Interative thesis combinging analyis, calcualtion and representation

February 3, 2020

Contents

1 Introduction

1.1 Motivation

In the beginning of my academical career I was aiming an educational degree i math and physics. While studying at the University of Tübingen I was also worked as a research assistant in the Institute of Physical Chemistry. In the course of my studies I was allowed to spend 4 month in a school to gain first experiences in teaching physics and math to students between 12 and 20 years. Even though it was a great experience and I enjoyed it a lot to bring new ideas and concepts to the students, I also felt a strong urge to continue learning and explore. At this time the possibilities I saw in focusing on scientific research and development seemed more attractive to me. My reasoning was, that I might be better for me do combine my preference for teaching and research rather at the university than a college. Long story short, after some years I ended up doing Phd. in the "Institute of Theoretical and Physical Chemistry" at the University of Tübingen. Lucky as I was I directly got the possibility to work in a project with an industrial partner. The focus was on building a new state of the art gas sensors. Besides the benefits of learning to regularly give

presentations and prepare reports, working in teams and having enough financial support for research activities, it also imposed some new problems needed to be solved. One of the biggest challenges was the continuous increase of sensor date in always shorter intervals. With the increasing industrialization and automatisation of sensor production and testing, the quantity high quality data increased dramatically. Luck as I was, I was not the only one facing problems with the increasing amount of data needed to be analyzed. In the beginning of each analysis, the specific task was not fully defined. It was rather an exploratory research to gain an idea about possible goals. In this case the the traditional way of creating software algorithms for well defined tasks, to take over the heavy lifting, did not fit too well. I rather had the need for efficient tools to reduce time consuming parts of the data analysis and a platform to combine these tools efficiently. So at this point of my career with my very basic knowledge in the area of programming and the need to solve urgent tasks, I was looking for an efficient solutions with a steep learning curve.

Until then my, and of some others in the lab, standard procedure of working with our data was mainly based on manual feature extraction and analysis. When Working with a limited amount of samples, manually doing these steps was acceptable and efficient. With the industrial cooperation gaining speed, the number of samples increased also rapidly. Soon we reached the point where the pre-processing of the data would easily consume a large amount of time. And that without even having started with the analysis. Luckily I was not the only one facing such problems and a project from the Python community gained more and more interest. Articles like [Unp14] pointed me to a new way of dealing with data and using the Python programming language. Regarding the fact, that Python was already a well established programming language, the introduction of the "IPtyhon notebook ecosystem" was making the use of Python for in an scientific work flow very attractive. As mentioned in this article: [Dav13]

> "... what they do offer is an environment for exploration, collaboration, and visualization."

I also realized the large potential for my working field. By learning Python I got efficient tools for calculations, analysis and representations. Additionally the new tools have been build specially with the focus on the goal to easy report result including the way they have been gained. The environment around the so called "Juypter notebooks" was the ideal piece, which I experienced as a missing block in the scientific work I was doing.

Besides my work for our industrial partner I also did fundamental research about semiconducting metaloxide gas sensors. Based on great research done before my time on gas sensors, my focus was now on numerical calculations of the gas sensors. Typically a theoretical model of a gas sensor is developed and the predictions are compared with experimental results. When publishing the results the peer review system assures, that the publications are well written, documented the on a solid and proven basis. But when reading such papers, I had the experience, that unfortunately many of them presented good results, but practical instructions on how to implement the presented models were often not given. The work of rebuilding the model and recalculating the results was therefore often not possible in a reasonable amount of time. This limitation can be a reason why a direct comparison model with experimental data from other sensors is often not done. It is also worth to mention, that in my experience the average experimental oriented researcher does not have the required programming knowledge to easily implement algorithm based on the presented work. But I am confident, that if an algorithm was given in an appropriate way, most research would benefit from the presented code.

Also with the results I gained while my Phd. thesis I was facing the same problem. Others under-

stood the value of my work but could not transfer it to their particular problem. At this point my graphical representation in the form of presentations, my detailed description and the corresponding algorithms of my work had been three separated parts. Facing this problem more and more, I decided not to ignore it anymore. I made the choice to try to combine representation, description and algorithm in one unified way. I decided to orient myself back to my "educational roots" and use the powerful ecosystem around the "Jupyter notebooks" to calculate, represent and describe the results for my thesis for this task. This is why my thesis is written with the large focus on suppling a introduction to this excellent toolbox.

On the one side this presented work will allow more people to gain insides in the sensing properties of semiconducting metal oxide sensors (SMOX). I hope that on the other side, this his thesis will also be a useful introduction into Python, specially IPython notebooks, for scientific work.

These hopes are not unfunded. In my time working at the "Eberhard Karls University of Tübingen" i was also assigned to give multiple lectures. One lecture branch has been the introduction to data mining. Often the lack of programming knowledge and the limited amount of lecture time did not allow a usage of Python as a supporting programming language. In these cases classical tools like Excel of Origin have been used to analyze example datasets for hidden facts. Most attendees have been very fast and understood the general concept of data mining. Just not being able to translate the general concept into machine understandable instructions stopped them from using them. When enough time was available, a short introduction into programming with Python gained a lot of interest and was generally seen as a positive experience. With just a short introduction already many advanced tasks can be performed, which would often even not possible with the "traditional" tools.

This thesis is therefore structured in such a way, that I will present my research results from the past years in a condensed form and additionally uses this opportunity to introduce and explain the importance of applying programming tools in the common work flow of scientific work. The potential of "outsourcing" repetitive tasks to machine executable scripts lies in the gain in investing more time in creativity and intelligence solutions. My hope is to bring with this thesis not only a deeper insight in the understanding of SMOX based gas sensors, but also help others to start a interesting journey into the wide area of data mining and machine learning with python.

I typically finish my introductions to Python by letting the students run their first commands. Typically for other introductions found around the globe, this is a program which outputs "Hello World.". For Python I prefer to execute some other commands with boils almost down to the philosophical essence on how "instructions" should be.

[1]:
```
import this
```

```
The Zen of Python, by Tim Peters

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
```

```
Although practicality beats purity.
Errors should never pass silently.
Unless explicitly silenced.
In the face of ambiguity, refuse the temptation to guess.
There should be one-- and preferably only one --obvious way to do it.
Although that way may not be obvious at first unless you're Dutch.
Now is better than never.
Although never is often better than *right* now.
If the implementation is hard to explain, it's a bad idea.
If the implementation is easy to explain, it may be a good idea.
Namespaces are one honking great idea -- let's do more of those!
```

1.2 SMOX based gas sensors

This section list a collection of high quality publications which cover all relevant information about semiconductor based gas sensors. This thesis is structured this way, that I will try to provide the required background in detail at the relevant places. So reading these publications now is not necessary to follow this thesis. This section is therefore a good point to come back if a more detailed view on the subject is desired. Especially when this thesis is used to apply the presented concepts to individual research topics.

When solving numerical problems it typical to introduction some assumptions, which are only valid under specific boundary conditions. These assumptions simplify the problem to a level where a numerical calculation is possible, but will reduce the validity of the results only to a small subset of all possible situations. The calculations in this thesis are also packed with assumptions and boundary conditions. My intention is to supply enough information to understand the relevance of the assumption and it's implications. With my presented work I do not claim to calculate all aspects of SMOX gas sensors, but present a tool which helps to understand specific aspects. The way of presenting this knowledge should lead/motivate others to adapt the presented work to individual other cases with completely different boundary conditions.

1.3 Jupyter Notebooks

"The Zen of Python" might not always be the primary directive of each developer, but the Python community consists most probably of many people how would consider the latter points as important. So did also the inventors of the IPython and Jupyter. A quick search will reveal multiple sources in the world wide web giving a detailed picture about what Notebooks are and how Jupyter in connected with them. Here I will not try to give an general overview about this tool and rather stick to the phrase "Learning by doing.". By explaining topic related parts this notebooks will guide an interested reader to the point where:

- understanding the fundamental instructions of Python
- using the basic functionality of the Notebooks
- fundamental understanding of SMOX based gas sensors

is gained.

It is worth mentioning, that the intention of such notebooks is to merge the essential tools of scientific work flows together. Data acquisition, preparation, analysis, representation and documentation all available in one place. The strength of not just sharing final conclusions in a nicely formatted way, but also being able to share the full stack of steps necessary to reach the final conclusion is essentially the strength of the Jupyter notebooks. This feature is already changing the way how scientific results are shared/published and was intentionally designed this way [RPGB17].

The default format of representing anything in a notebook in based on "Markdown". Wikipedia summarizes Markdown like this:

> Markdown is a lightweight markup language with plain text formatting syntax. Wikipedia

This means a document is formatted by writing plaintext and special text blocks are interpreted as formatting commands. E.g. **BOLD** letters are generated by encapsulating the text with `** Text here **`, headings are generated by starting the heading with `#`. Depending on the number of `#`, subsections are created. I will not go into detail here about the features of Markdown. Many features are used in this notebook and are directly accessible by double-clicking the text element. The plain text will reveal the way it was created, and the execution of the cell with `CTRL+ENTER` will reveal its Markdown formatted representation. One other handy feature of the Jupyter ecosystem I use is the ability to transform notebooks into multiple other formats. Just to name some: HTML, WORD (DOC, DOCX), Latex. The tool `nbconvert` is used internally to convert the Markdown formatted representation into other formats. For this thesis the default option "Export as PDF" under the File option generates a Latex based PDF file. Mastering-markdown is a web page, where I found some hints on how to format my notebook. For instance I gained the ability to make block quotes from this page based on this example:

As Kanye West said:

> We're living the future so the present is our past.

To learn how to use notebooks it is best to use them in an interactive environment. The next section will explain how to obtain one for free!

1.3.1 Installation guide

The easiest way to get started would be to use the Anaconda distribution. Anaconda bundles multiple different tools and installs them in the operation system. Anaconda will take care of cross dependencies and handle the update process of the software. This is not the only way to get started with "Jupyter Notebooks" but surely an very fast and easy one. HERE is additionally a presentation I use for my lectures to guide students into the world of Python and here one example of it's usage.

1.3.2 Example Notebook - Sneak preview

Besides the first example `import this`, here is a very basic example which should prepare exited reader on whats coming next.

"Simple is better than complex." The Jupyter environment is equipped with "magic commands", which are not part of the base programming language (i.e. Python), but rather a helper instruction to simplify common tasks. Magic commands always start with `%` and are followed with an instruction. I will demonstrate in this example the use of the `%pylab inline` instruction. This modifies the current programming space to become a lab nicely equipped for scientific work tasks. For instance a chemistry lab is commonly equipped with a balance, a water tab and a fire extinguisher, and in this case a "pylab" is (besindes many others) equipped with a data handing, a plotting and a calculation tool. The additional parameter `inline` makes sure, that the figures will be along with this document. So let's setup a "pylab" and run some lines of code. (The plotting is handled in the background by Matplotlib [Hun07])

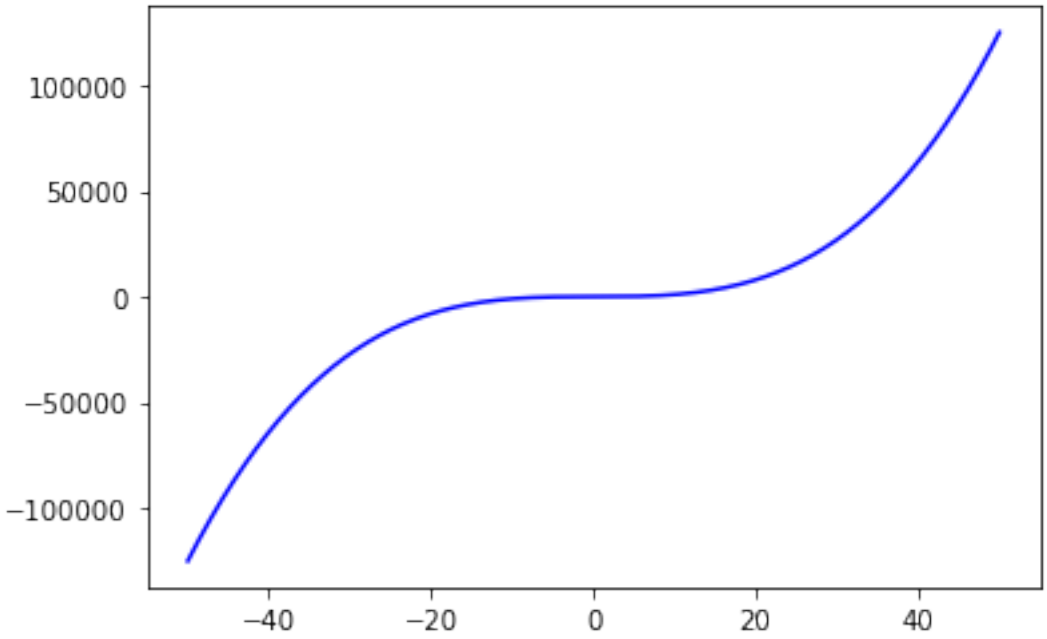
[5]:
```
%pylab inline

#get a list of 5000 points between -50 and 50
xs = linspace(start=-50, stop=50, num=5000)

# Python way of adding a comment is the '#' character
# Python way of writer 'x to the power of y' is x**y
# here I am calculating thethe third power of x for 5000 points between -50 and␣
 ↪50
ys = xs**3

#plot it
fig = plt.plot(x,y,'b')
```

```
Populating the interactive namespace from numpy and matplotlib
```



The `linspace` function generates a 5000 linear distributed points in the interval [-50,50) and saves those points in the xs variable. y is just the third power of each point. `plt.` is a submodule defined by the magic command `%pylab` which handels data plotting in a very simple way. `plt.plot(x,y,'r-')` for example plots x vs. y with a red line. The thesis is done in the notebook to offer the reader the possibility to directly work with newly gained knowledge. Therefore the upper block is a good opportunity to do the fist steps in Python. For instance change the line format to 'b' (blue). Or to 'b-.'. Change the exponent from 3 to 3.1. To do so:

- click on the code box
- edit the field
- Add the following line `plt.title('The power law')`
- go back with CTRL-Z to correct your mistakes
- hit CTRL-ENTER to execute the code

2 Conclusion

Since the motivation for this "interactive" thesis should be clear now, I would like to come in the next section now to my actual research topic: "Numerical calculation of semiconducting metal oxide (SMOX) based gas sensors". In the next chapter I will demonstrate how theoretical numerical calculations are used to for chemical sensors are used to better understand experimental results. In the following chapter I will demonstrate how such knowledge could be used to improve the performance of a sensor regarding it's sensitivity and selectivity.

Follow this link to come to the next section.

3 About the PDF-Version of this work

This notebook was not intended to be used as a printed hard copy or as a PDF. The provided PDF servers just as an low level representation of the original work. It should give potentially interested readers an easy entry point to Python (or any other programming language) supported science. Many of the implemented features in these notebooks like interactive widgets, animated data representations and live code examples will not work in the PDF-version. Only a static snapshot of the mutable representation can be represented outside the notebook, at best. The benefits of interacting with the presented work in a notebook should motivate the reader to use the notebook.

Nevertheless the integrated function to export a notebook to a latex based version offers a very nice way to publish results in a printable way. So please keep in mind, that the PDF-verison may not be able to represent all the features of the notebook as intended and some links might not work as expected. You are strongly encouraged to switch to the Jupyter presentation of this work and experience the full potential of such notebooks.

3.1 Equations

In a typical scientific thesis and textbook, relevant equations are referred by numbers of identifiers. In Latex this is done, by assigning a label and a tag to an equation. If this equation needs to be referred to, a pointer to the reference is added, and the tag is used for the representation of this equation. As an example, an arbitrary equation from this thesis is used. The (internal) reference of this equation is 'second_derivative', while the printed representation (tag) is "Second derivative".

$$\frac{dV^*}{dr^{*2}} = 1 - n^*(V^*) - \frac{2}{r^*}\frac{dV^*}{dr^*} \tag{Second derivative}$$

To refer to this equation a reference can be added which will look like this (Second derivative). The underlying mechanisms will link the reference with the equation, print the label, and add an hyperlink, to be able to jump to the equation. This feature works either in notebook or PDF representations. One drawback of the notebook representation is, that the equation and reference need to be in the same code block/cell. Otherwise the reference is not working. Instead of linking the equation correctly, ??? will be represented. My opinion is, that this will change will evolving improvements of Jupyter is is just a temporal flaw. Since the PDF version is processed by a Latex interpreter in total, the "one code block" limitation does not exist there. To demonstrate this, I will try to reference the eqaution in the next code block.

Here the same reference to the equation: Second derivative. In the PDF-Version this will work correctly and in the Notebook-Version only ??? should be visible (until now).

3.2 Tables

With the numerical calculations performed in this thesis, data will need to be represented also. One way of doing this is by plotting the the data. This is also suitable for a static PDF-Version of this thesis, which often is useful/necessary to have. This feature is well implemented and the transfer from Notebook to PDF version works well. Another way to display data are tables. Similar to the well known, omnipresent tool for working with tables, Excel, the Python universe has its own, but similar, tool. It's called Pandas. As a primer I will give here a very short introduction

to Pandas. What sheets are for Excel, are Dataframes for Pandas. Here a simple example how to create a Dataframe in analogy to the previous programming example:

[6]:
```
%pylab
import pandas

#create some x values
x = linspace(start=0,stop=5,num = 10)

#the y values will the the square of the x values
y = x**2

#Put them in a Dataframe and reference this new Dataframe with the variable `dF`
dF = pandas.DataFrame({'x':x, 'y':y})
```

```
Using matplotlib backend: Qt5Agg
Populating the interactive namespace from numpy and matplotlib
```

This simple example of a Dataframe should demonstrate the basic concept of Dataframes. Once data is in the `DataFrame` format, the are infinit ways to transform/slice/merge/... it, to bring it into the desired shape. Along this thesis, some of the functionalities of Dataframes will be used and explained. Since the Jupyter environment is still "under construction", the transfer from Juypter to PDF for notebooks does not work to well until now. This does not mean, that is is not possible, it is just not yet implemented in the default/vanilla environment. This is way I want to highlight the small tweak that is at this point still needed to have a comparable output in Juypter notebooks and printed PDFs. To bypass this obstacle the default behavior of pandas needs to be altered (similar to this post: Latex-Tables Monkey Patch). This following patch brings DataFrames in the appropriate shape to by nicely represented in both representations.

3.2.1 Before the patch

[7]:
```
display(dF)
```

	x	y
0	0.000000	0.000000
1	0.555556	0.308642
2	1.111111	1.234568
3	1.666667	2.777778
4	2.222222	4.938272
5	2.777778	7.716049
6	3.333333	11.111111
7	3.888889	15.123457
8	4.444444	19.753086
9	5.000000	25.000000

3.2.2 The patch

[8]:
```
import pandas
pandas.set_option('display.notebook_repr_html', True)

pandas.set_option('display.notebook_repr_html', True)

def _repr_latex_(self):
    return r"""
    \begin{center}
    {%s}
    \end{center}
    """ % self.to_latex()

pandas.DataFrame._repr_latex_ = _repr_latex_  # monkey patch pandas DataFrame
```

3.2.3 Prettier tables after the patch

[9]:
```
display(dF)
#"Test Tabel"
```

	x	y
0	0.000000	0.000000
1	0.555556	0.308642
2	1.111111	1.234568
3	1.666667	2.777778
4	2.222222	4.938272
5	2.777778	7.716049
6	3.333333	11.111111
7	3.888889	15.123457
8	4.444444	19.753086
9	5.000000	25.000000

In Jupyter notebooks the output will look very similar. In the PDF version of this thesis, those two output will differ(, unless a newer version of Jupyter fixed this issue.) To have this thesis in a PDF printable form, I will use the presented patch to unify the output. On the other side I suggest not to use this patch, and rather work with the Jupyter notebooks and its default representation, since the patch has also its downsides. But this would go beyond the scope of this remark about the patch.

4 Bibliography section

References

[Dav13] DAVENPORT, Thomas H.: *The Rise of Data Discovery*. https://www.datanami.com/2016/05/04/rise-data-science-notebooks/. Version: 2013

[Hun07] HUNTER, John D.: Matplotlib: A 2D graphics environment. In: *Computing in Science and Engineering* 9 (2007), may, Nr. 3, 99–104. http://dx.doi.org/10.1109/MCSE.2007.55. – DOI 10.1109/MCSE.2007.55. – ISSN 15219615

[RPGB17] RANDLES, Bernadette M. ; PASQUETTO, Irene V. ; GOLSHAN, Milena S. ; BORGMAN, Christine L.: Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study. In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (2017). http://dx.doi.org/10.1109/JCDL.2017.7991618. – DOI 10.1109/JCDL.2017.7991618. – ISBN 9781538638613

[Unp14] UNPINGCO, José: *Python for signal processing: Featuring IPython notebooks*. Bd. 9783319013. Cham : Springer International Publishing, 2014. – 1–128 S. http://dx.doi.org/10.1007/978-3-319-01342-8. http://dx.doi.org/10.1007/978-3-319-01342-8. – ISBN 9783319013428