

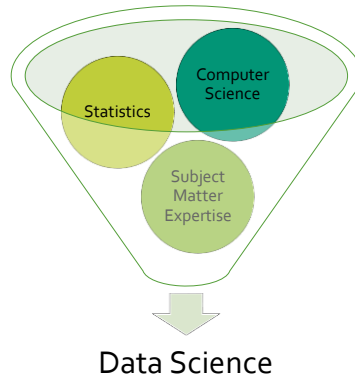
DATA SCIENCE

An Introduction to Computer Science



Let's learn about Data Science.

Data Science

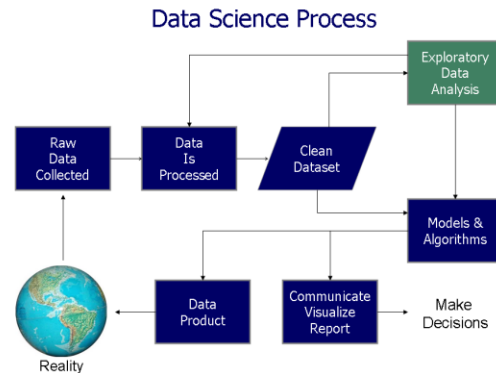


As previously mentioned, Data Scientists use data and programming to answer real-world questions.

Although different from the field of Computer Science, there are many overlaps in the skillsets needed.

Both careers are extremely popular right now, and it is valuable to learn both.

Workflow

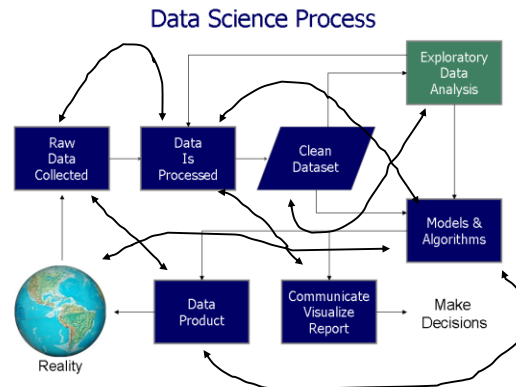


O'Neil, Cathy and, Schutt, Rachel (2014) Doing Data Science,
O'Reilly ISBN: 978-1-449-35865-5.



The diagram shown here models one way to conduct a Data Science analysis. In general, data is collected from the real world, and then processed into a suitable format. The data scientist then iteratively explores the dataset, refining it further, and developing research questions. They answer these questions by creating plots, running statistical analyses, and making models from the data. The results that they gather are reported back to interested stakeholders, who make decisions that affect the world. Of course, in practice, each data scientist develops their own process, but this model is well-regarded.

Non-linear Workflow

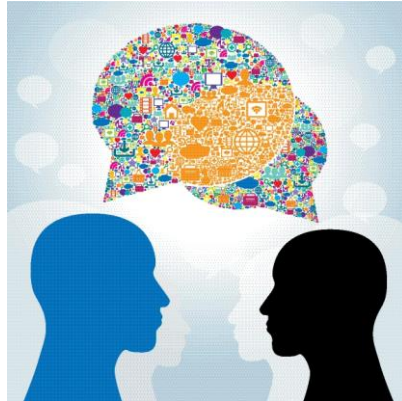


O'Neil, Cathy and, Schutt, Rachel (2014) Doing Data Science,
O'Reilly ISBN: 978-1-449-35865-5.



Data scientists rarely work straight from beginning to end. Although the workflow shown may suggest an orderly sequence of events, the reality is that data scientists move from phase to phase as needed. Sometimes, you need to revise your questions after you find your first answers. Other times, you realize you need a different dataset in order to answer your questions

Telling a Story



<https://blogs.sas.com/content/sascom/files/2015/04/470043831.jpg>



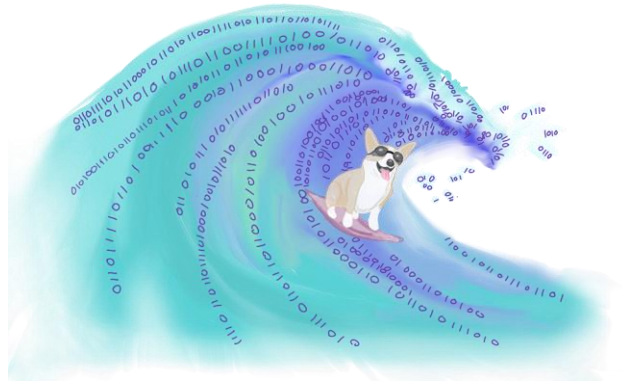
In many ways, a good scientist becomes a good storyteller.
You are collecting data and analyzing it in order to tell a story.
We very rarely learn universal truths, but are instead building up evidence to support a particular hypothesis.
Keep your audience in mind, and the story that you ultimately want to tell.

Finding Data



We simultaneously live in a data-rich world and a data-poor world. More and more processes and systems, both human and computational, create data. However, this data is often kept under lock and key to protect individuals, corporations, or governments. Further, many potential sources of data are not collected for pragmatic reasons. You may find that you want a particular dataset, but cannot get access to it. Other times, you will be given a tidal wave of data and you will struggle to deal with the scale.

A Tidal Wave of Data



More and more data is available in the world each day.

This tidal wave of data has led to the term "Big Data", which can refer to data is high in volume, changes rapidly, or has a very complex structure.

Of course, the amusing secret is that most data is not big, but that does not mean these smaller datasets are not useful.

You may eventually learn computational techniques to process big data, but for now you should still appreciate the power of small datasets.

Advanced Data Science

Areas

- Machine Learning
- Data Wrangling
- Data mining
- Graph Analysis
- ...

Tools

- Pandas
- Scipy
- Numpy
- Scikit-learn
- ...



There are many topics in Data Science that we do not have time to cover.

Although it can be tricky to learn how to use advanced techniques in areas like Machine Learning, you might be surprised by what you can accomplish.

You are encouraged to continue learning more about these advanced techniques and tools.

For now, focus on the basics: making questions and building answers using basic data processing.