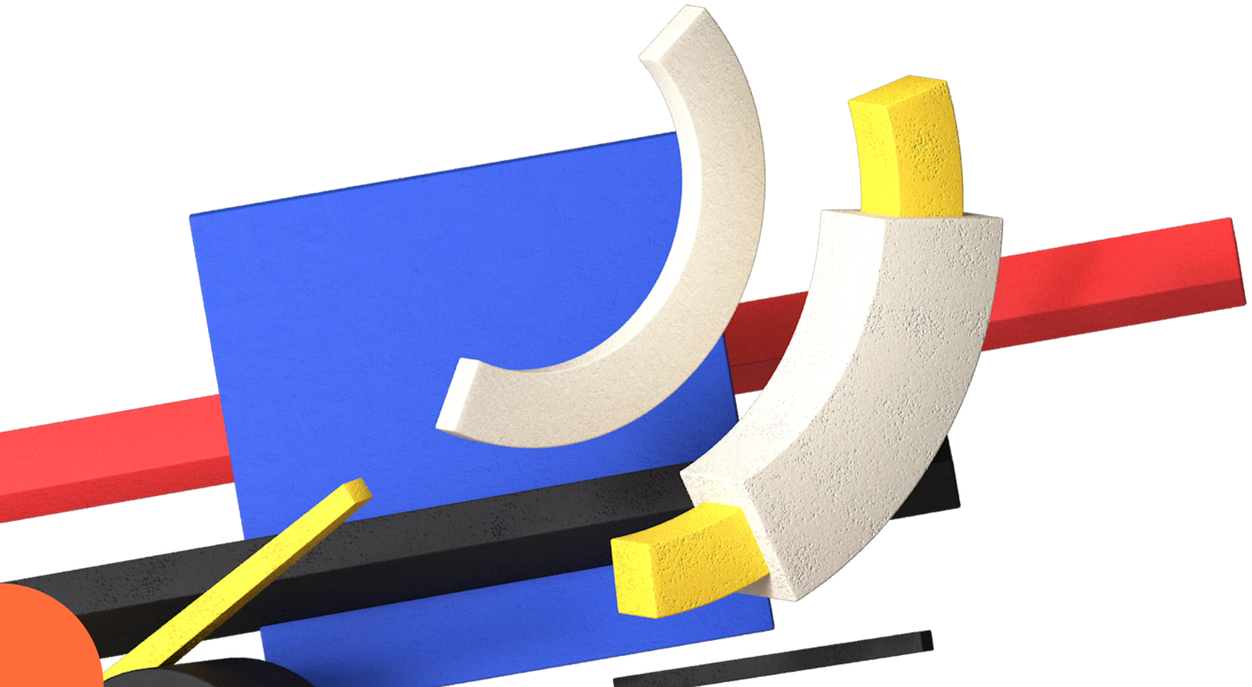
JAWAHARLAL NEHRU UNIVERSITY

Data Science and Statistics

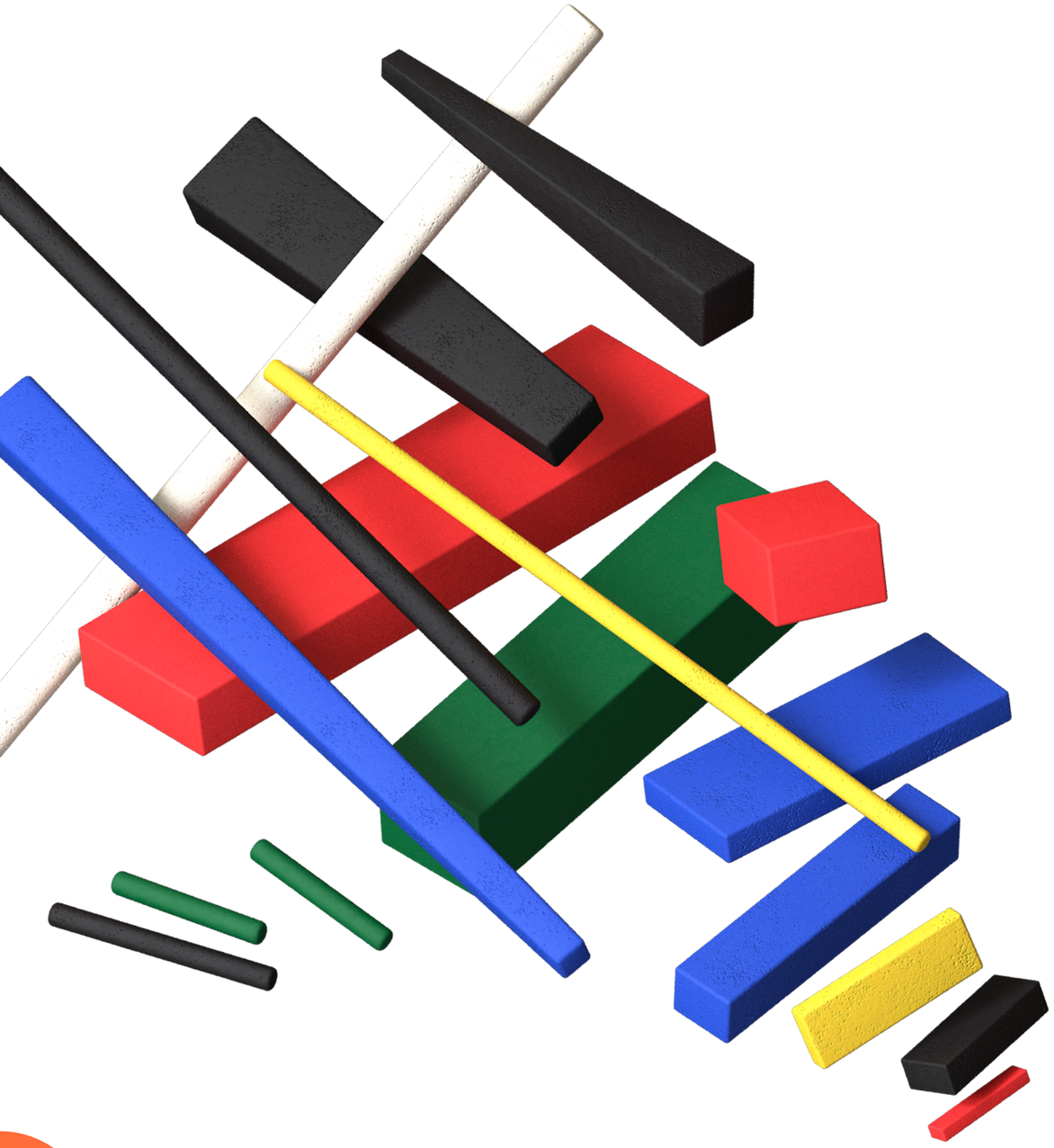
A workshop by C.o.L.D, the Computer Science Club of School of Engineering

Agenda

What's in store



- “Data! Data! Data!” he cried impatiently. “I can’t make bricks without clay.”
- The Zen of Python. And maybe more python.
- Is there anything more useless or less useful than Algebra?
- Facts are stubborn, but statistics are more pliable.
- To write it, it took three months; to conceive it, three minutes; to collect the data in it, all my life.
- Visualization gives you answers to questions you didn’t know you had.
- Art and a Science - Working with Data
- Data Science Afterthoughts
- Data science is the discipline of making data useful.
- And now, once again, I bid my neophyte progeny go forth and prosper.



Some Instructions

- The presentation and Jupyter notebooks used in this workshop will be available for download from our website after the workshop.

- Invite link of our Discord server would be live on our website during the workshop. Anyone interested can join it during that time.

- We have uploaded a cheat sheet for this workshop on our website. Visit our site and download it for reference during the workshop.

Note

- This is not a teach and learn session.
- Illuminating but impractical examples.
- Libraries stop you from understanding data science.

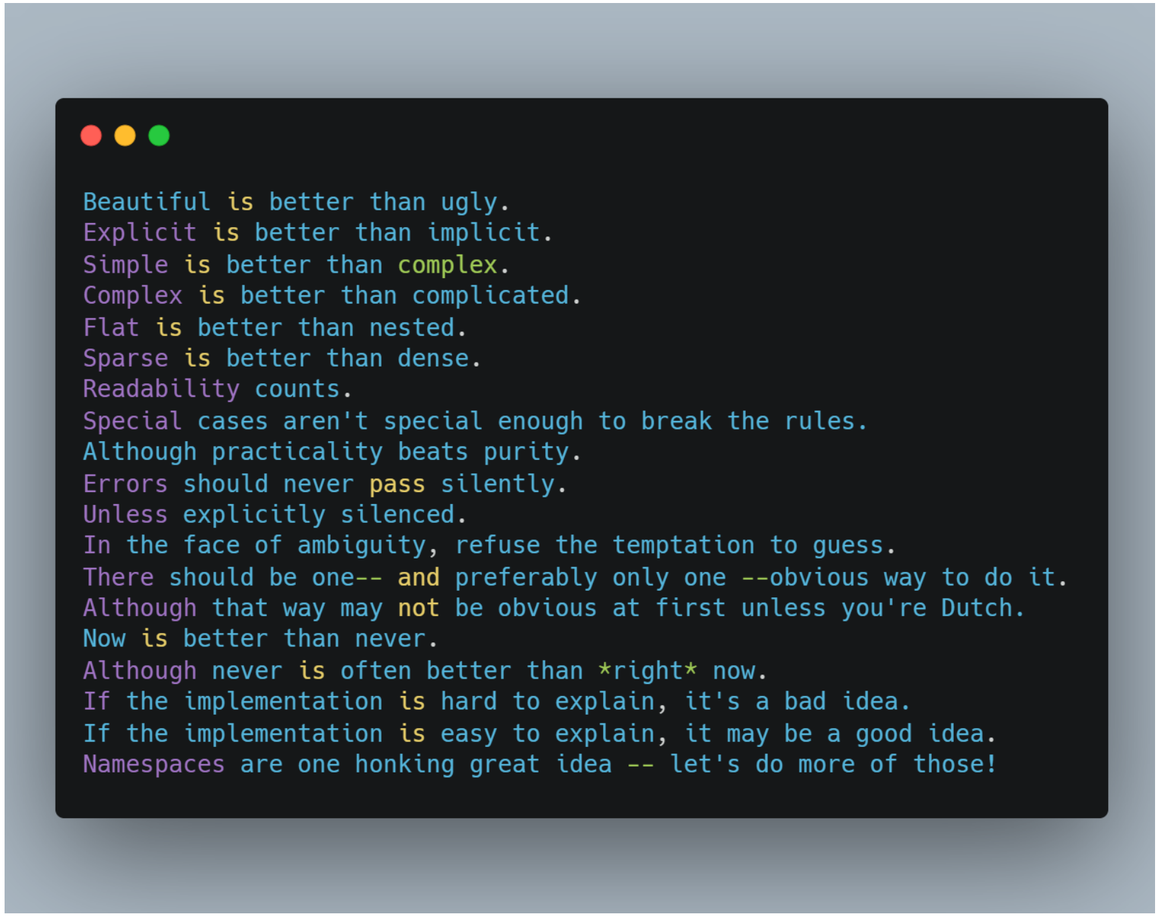
“Information is the oil of the 21st century, and analytics is the combustion engine.

Data science is the art and science of programitcally analyzing a dataset.

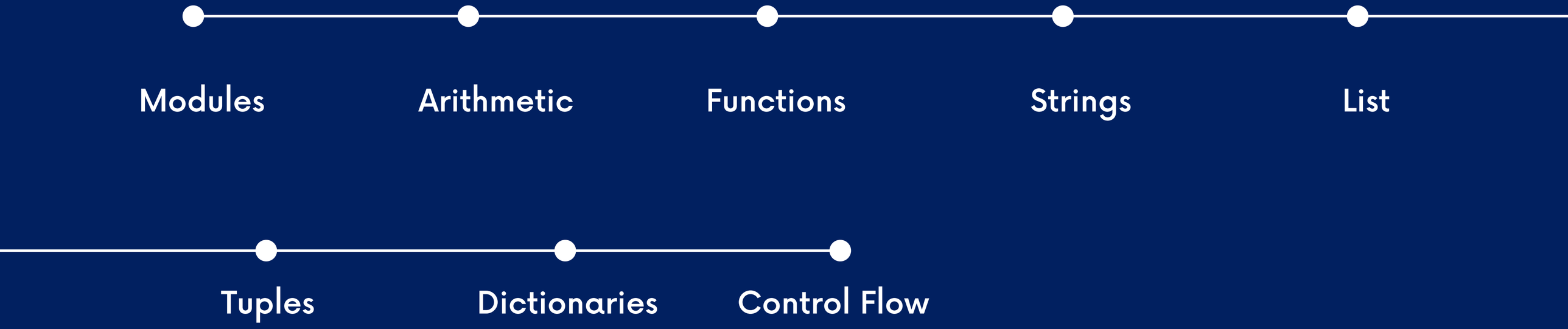
The future is defined by data science.

Data science practitioners are some of the most highly paid employees in the industry commanding an average salary of 17LPA in India and many times that abroad.

The Zen of Python

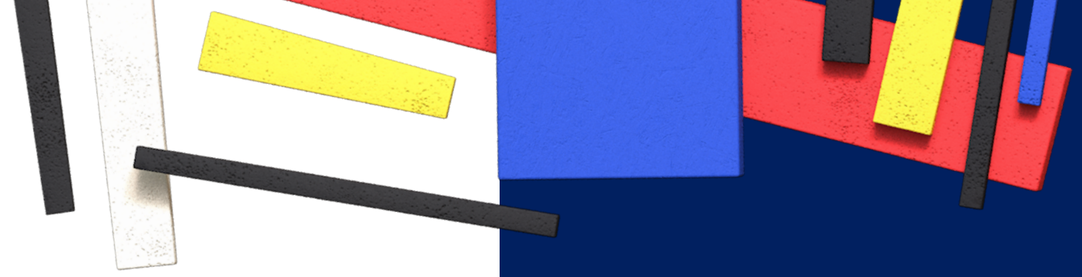


And maybe more Python.



Linear Algebra

Is there anything more useless (/s) or less useful than Algebra?



Vectors

- Abstractly, vectors are objects that can be added together (to form new vectors) and that can be multiplied by scalars (i.e., numbers), also to form new vectors.
- Concretely (for us), vectors are points in some finite-dimensional space. Although you might not think of your data as vectors, they are a good way to represent numeric data.

Matrices

- A matrix is a two-dimensional collection of numbers. In Python, we represent matrices as lists of lists, with each inner list having the same size and representing a row of the matrix.

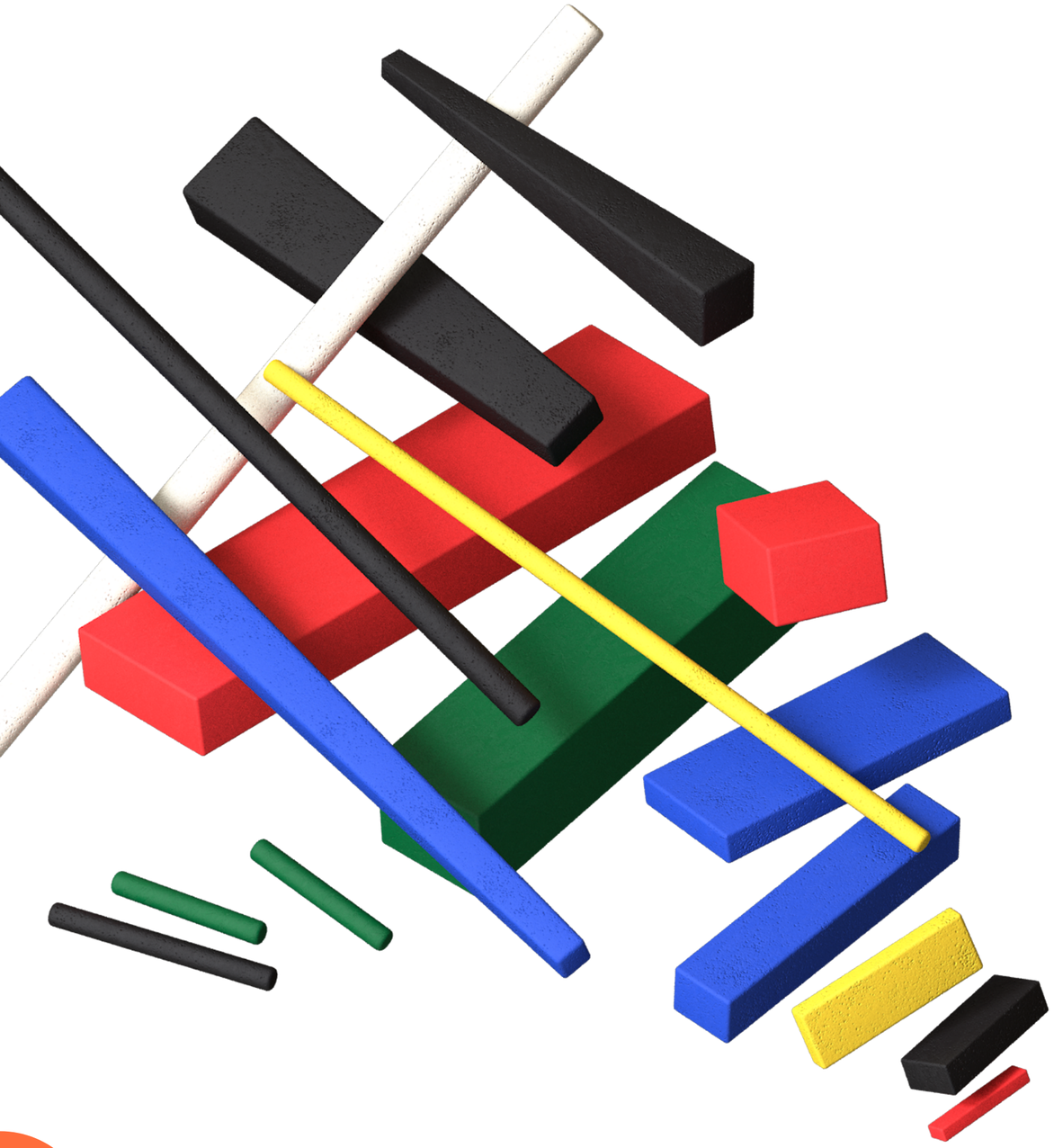
For Further Exploration

Linear algebra is widely used by data scientists (frequently implicitly, and not infrequently by people who don't understand it). It wouldn't be a bad idea to read a textbook.

You can find several freely available online:

- Linear Algebra, from UC Davis
- Linear Algebra, from Saint Michael's College

If you are feeling adventurous, Linear Algebra Done Wrong is a more advanced introduction

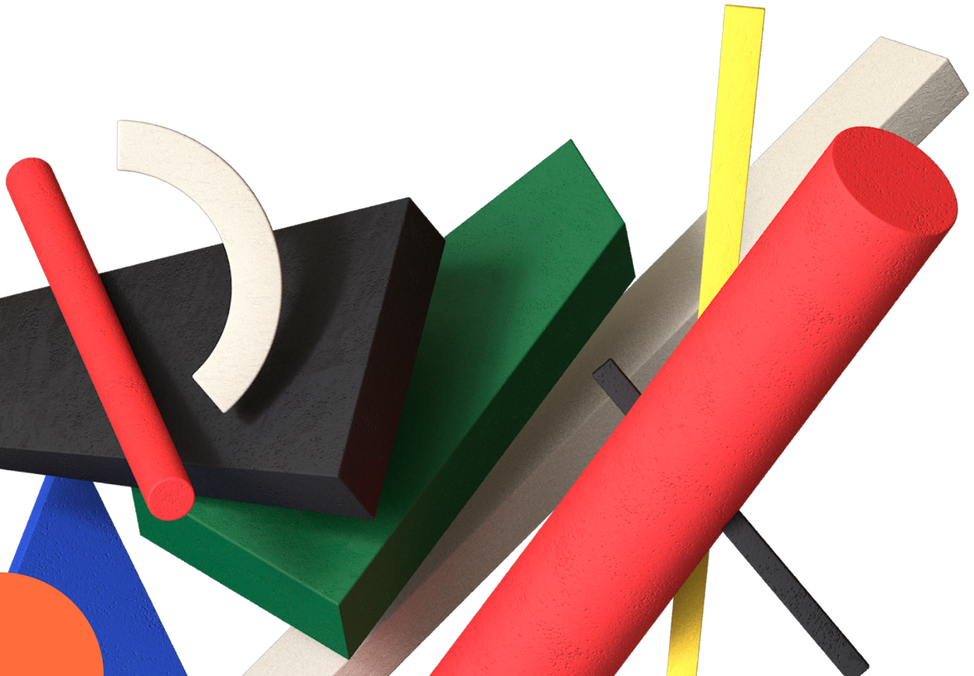


Statistics

Statistics refers to techniques with which we understand data. It is a rich, enormous field, more suited to a shelf (or room) in a library rather than a slide in a presentation, and so our discussion will necessarily not be a deep one.

Instead, we'll introduce two important concepts which make you just enough to be dangerous, and pique your interest just enough that you'll go off and learn more.

Statistics using case studies



Case Study 1: Correlation

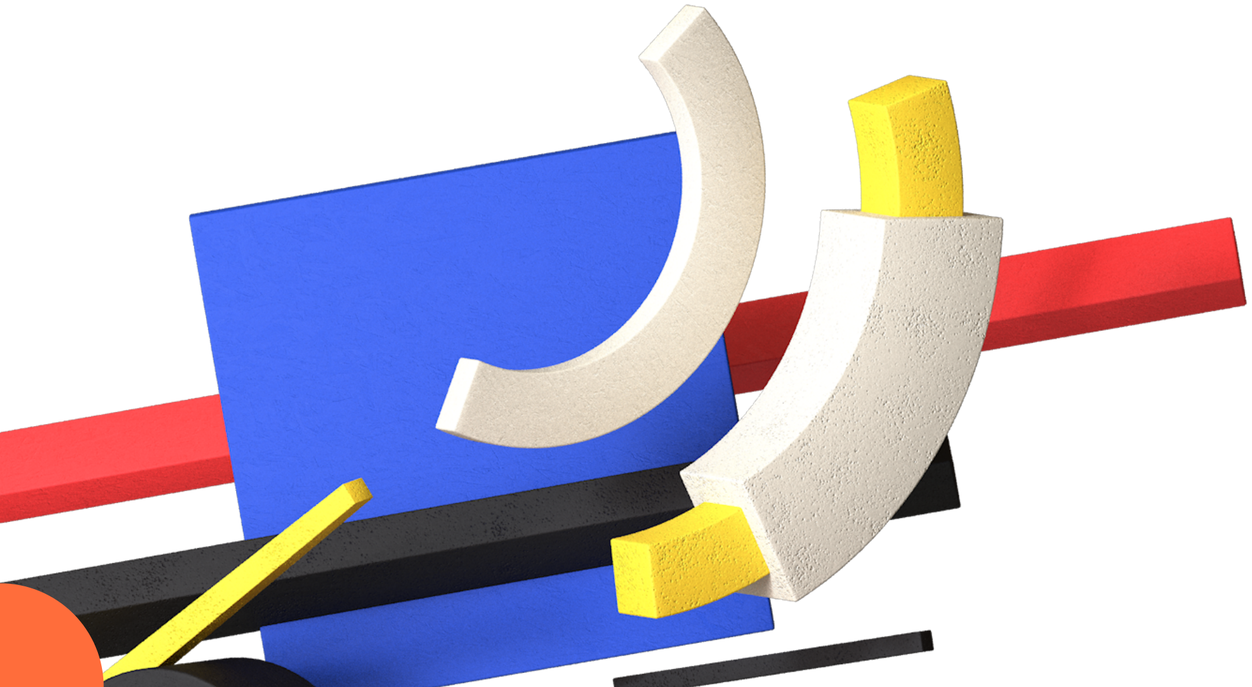
DataSciencester's VP of Growth has a theory that the amount of time people spend on the site is related to the number of friends they have on the site (she's not a VP for nothing), and she's asked you to verify this. We use correlation to test her hypothesis.

Case study 2: The physics lab

You are doing an experiment in the Physics lab and you obtained some readings. Now learn how to obtain a function that approximates your readings the best.

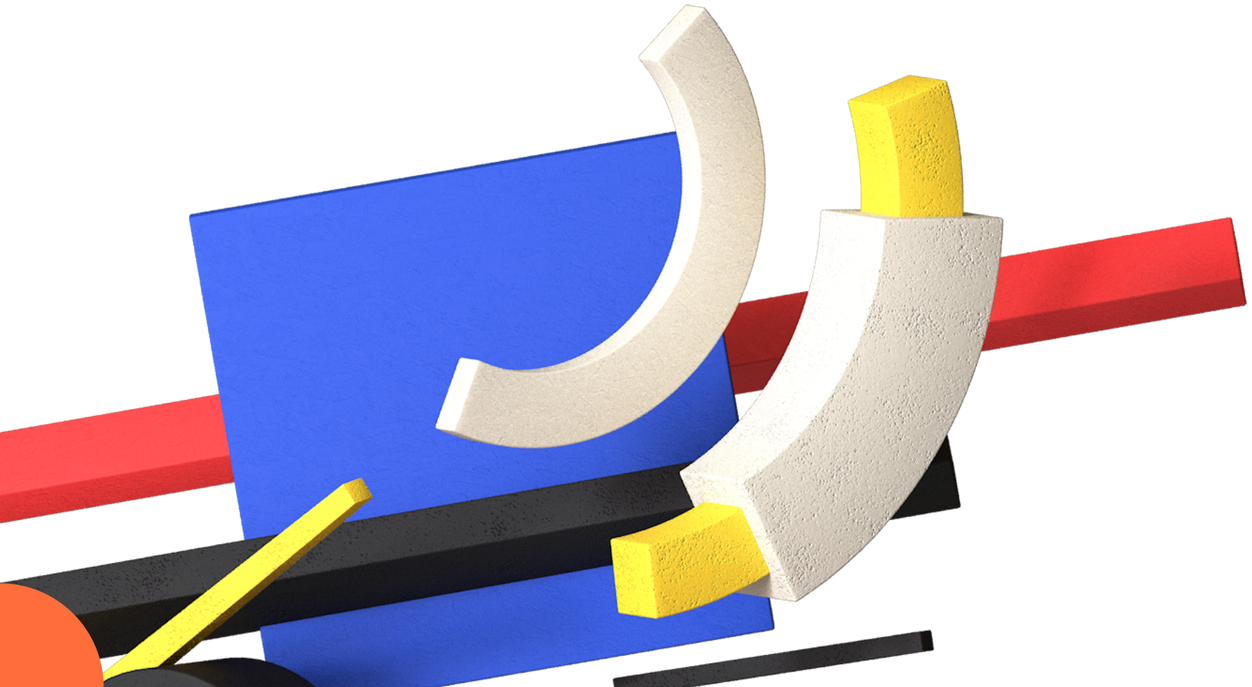
Case Study 1

Correlation



- After digging through logs you come up with two lists. You sort the lists and clean the data so num_friend and daily_minutes correspond with each other.
- Now lets move to the Jupyter notebook to analyze it further and understand how correlation can be used to prove or disprove a claim.

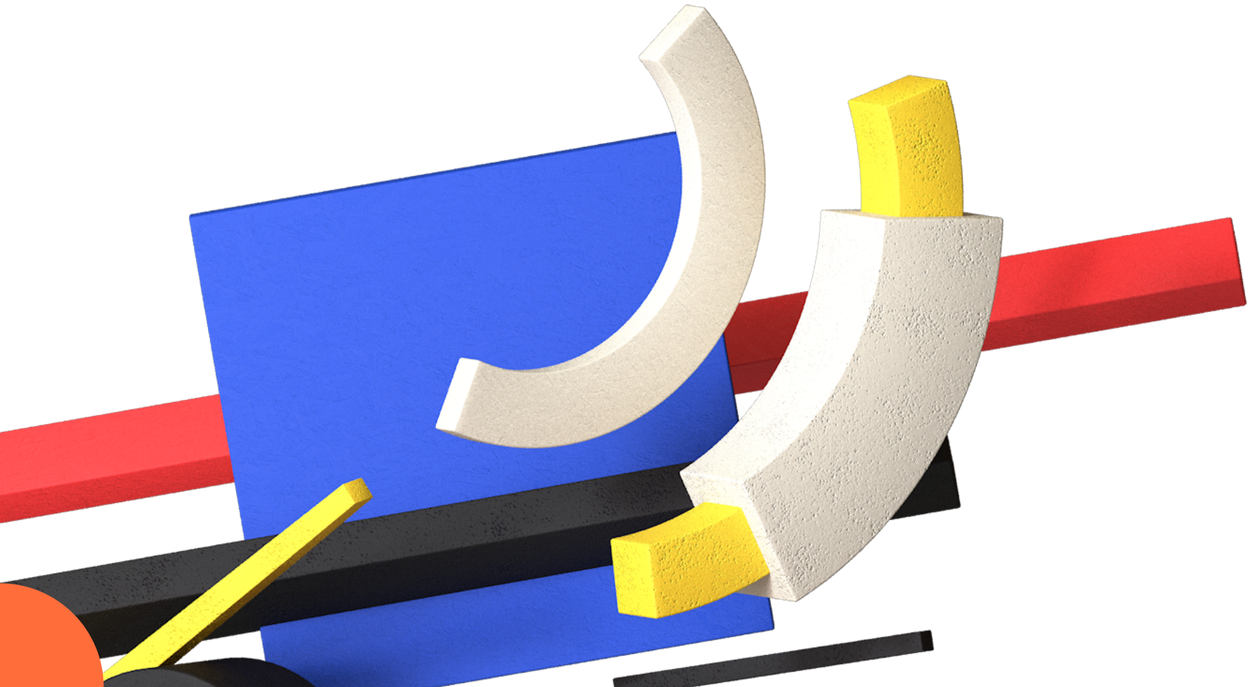
Conclusion



- The claim of the VP looks to be correct!
- However remember causation vs correlation.

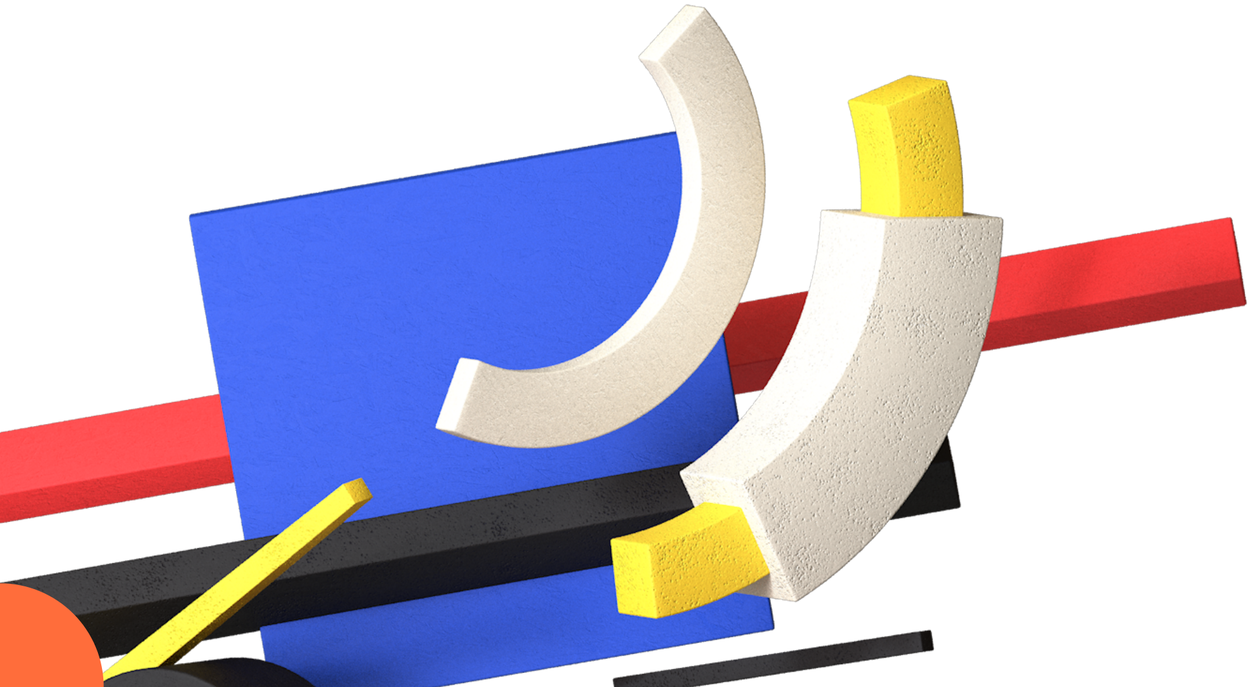
Case Study 2

Curve Fitting



- Curve fitting is a method to fit a test function to a set of observations.
- It has great use in scientific and data science applications.
- Here a set of observations for a capacitor are taken. How will we plot a graph that closely approximates the observations?

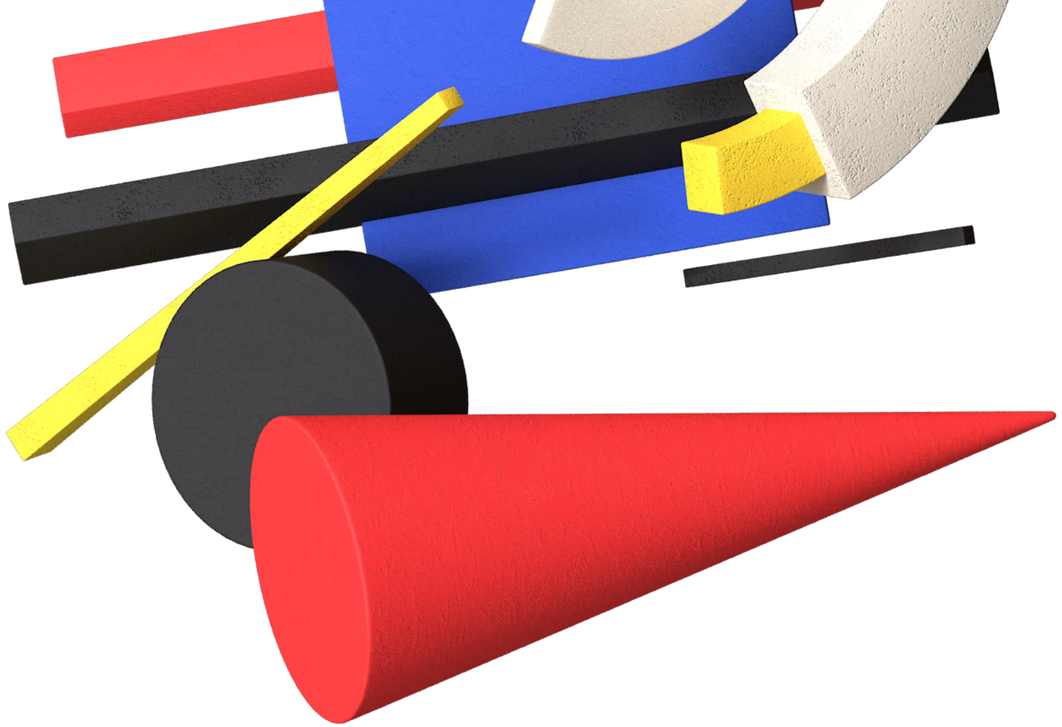
Conclusion



- The best curve is approximated by the scipy optimize function.
- The goodness of fit can be tested using the chi square test.
- This technique can be used for various practical applications.

To write it, it took three months; to conceive it, three minutes; to collect the data in it, all my life.

Obtaining Data



Reading Files

Data can be read from files directly using Python.

Web Scraping

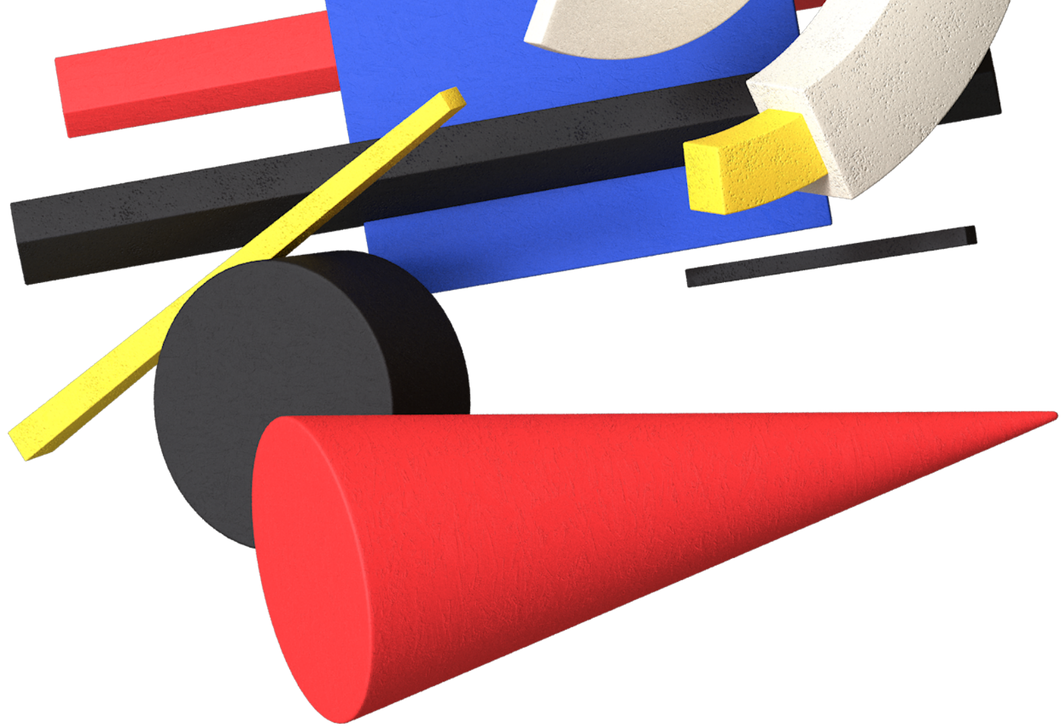
Data can be obtained from web pages using Python libraries like requests and lxml.

Querying APIs

APIs allow you to explicitly request data in a structured format. This saves you the trouble of having to scrape them!

Obtaining Data

Further Exploration



pandas

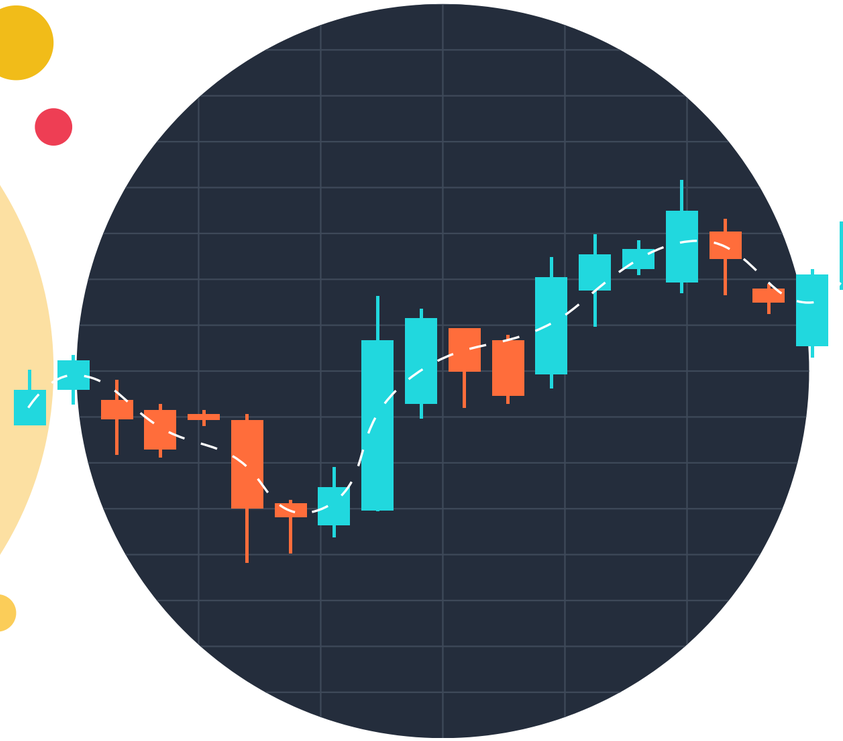
pandas is the primary library that data science types use for working with (and, in particular, importing) data.

scrapy

Scrapy is a more full-featured library for building more complicated web scrapers that
do things like follow unknown links.

Visualization gives you answers to questions you didn't know you had.

Visualising Data with Matplotlib



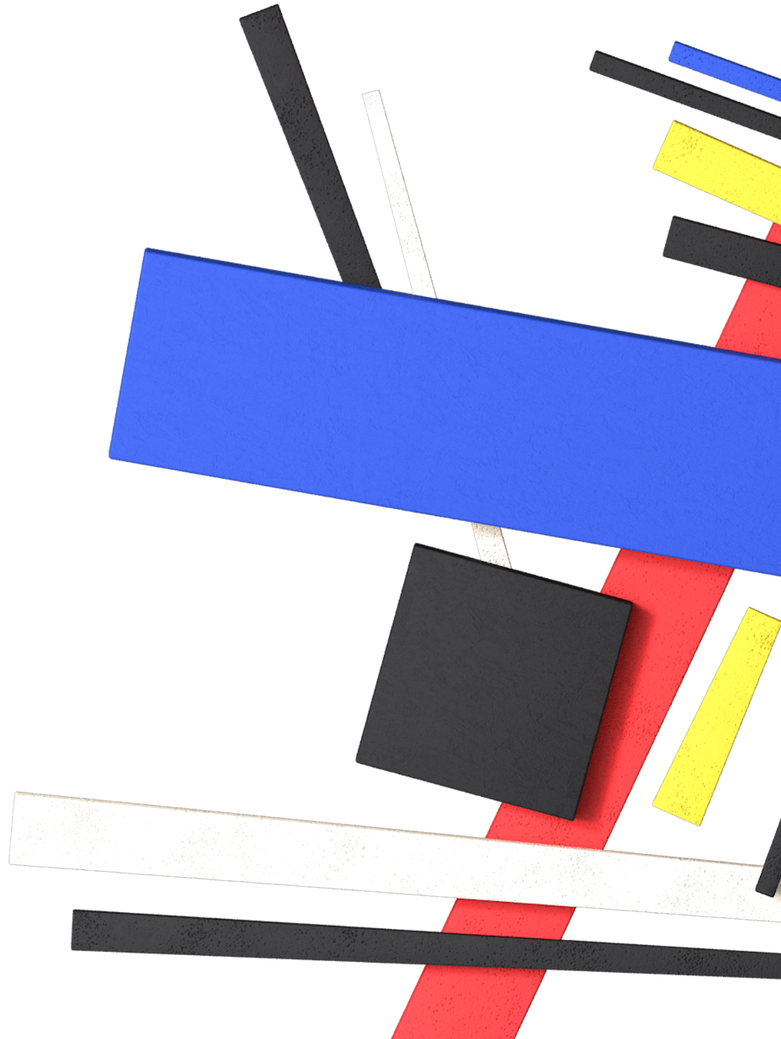
Seeing is believing...or is it?

A fundamental part of the data scientist's toolkit is data visualization.

Although it is very easy to create visualizations, it's much harder to produce good ones.

There are two primary uses for data visualization:

- To explore data
- To communicate data



Lets try out some graphs in jupyter notebook

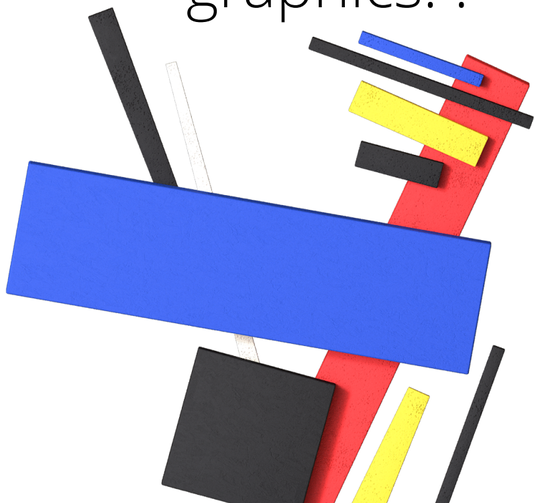
For further exploration...

Seaborn is built on top of matplotlib and allows you to easily produce prettier (and more complex) visualizations.

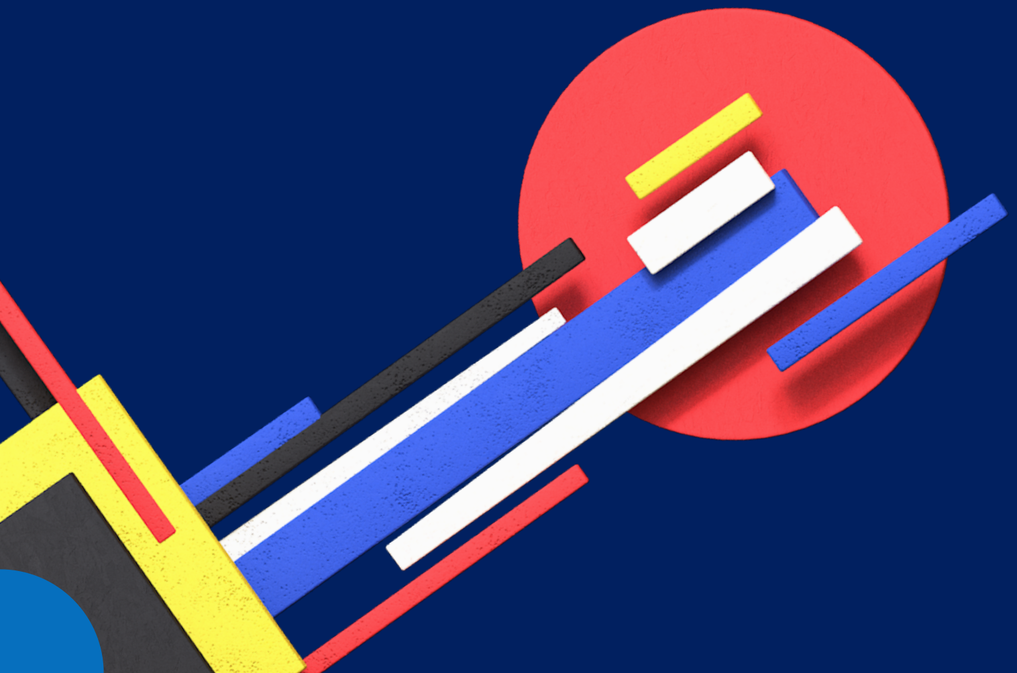
D3.js is a JavaScript library for producing sophisticated interactive visualizations for the web. Although it is not in Python, it is both trendy and widely used, and it is well worth your while to be familiar with it.

Bokeh is a newer library that brings D3-style visualizations into Python.

ggplot is a Python port of the popular R library ggplot2, which is widely used for creating “publication quality” charts and graphics. .



Data Science - The Afterthoughts



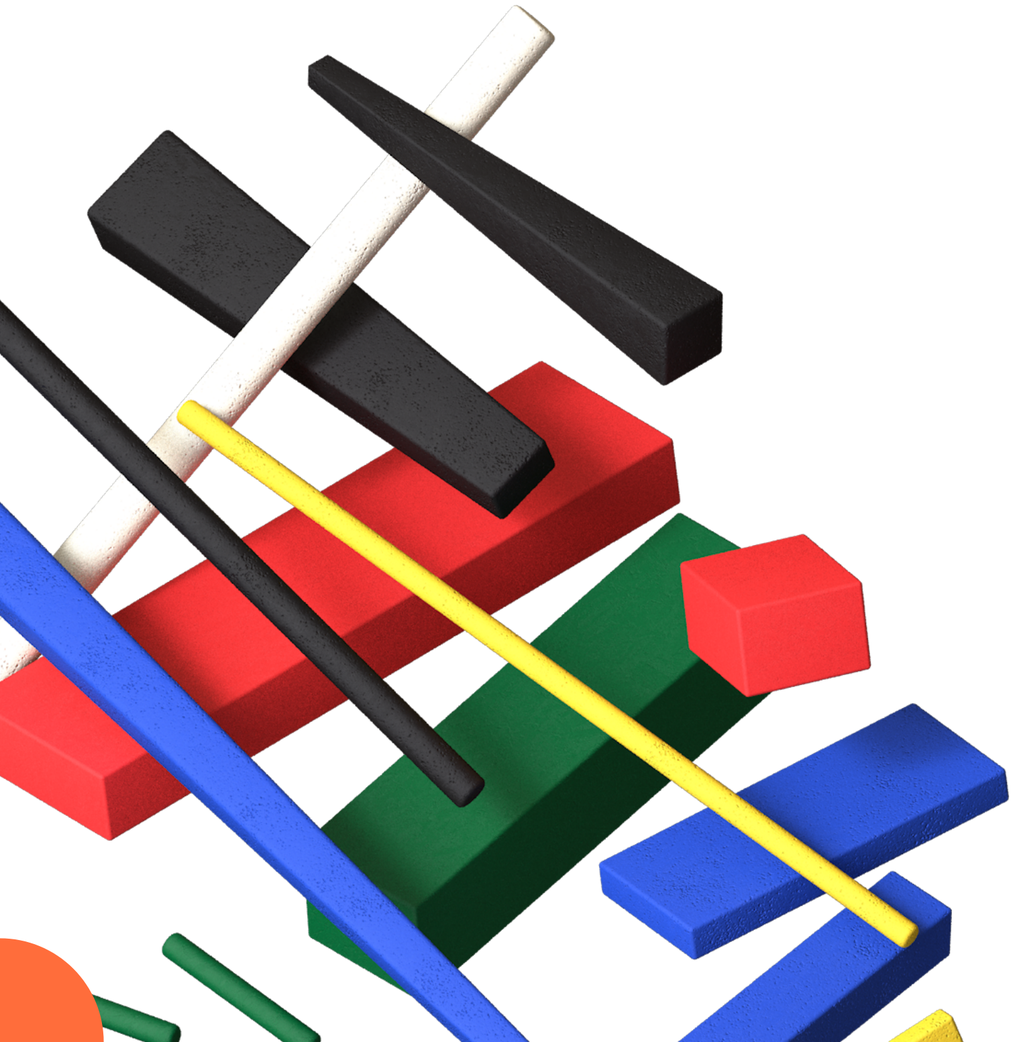
1. I am always ready to learn although I do not always like being taught.

2. Black boxes. Black black boxes!

Data science is the discipline of making data useful.

Applications of Data Science

The world is one big data problem



Data Science is everywhere

Let us discuss a few applications of data science in real world applications.

Lets find out why mega corps like Facebook, Google, Amazon etc spend millions on their data science departments.

Recommender Systems

O nature, nature, why art thou so dishonest, as ever to send men with these false recommendations into the world!

Another common data problem is producing recommendations of some sort.

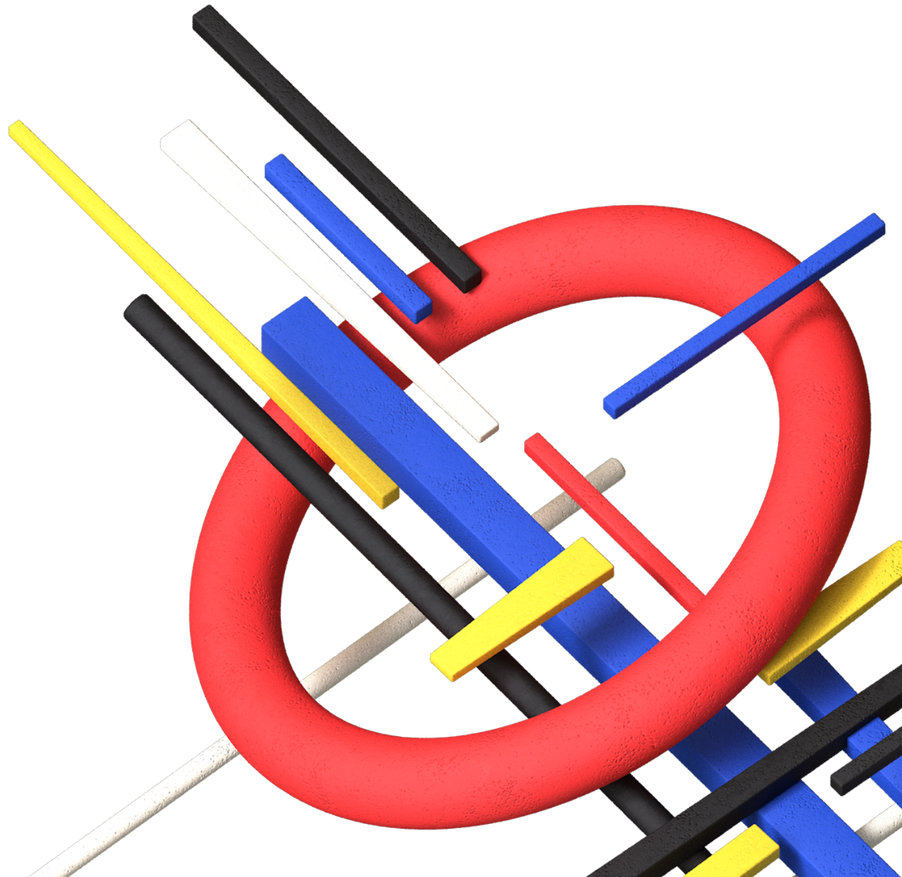
Netflix recommends movies you might want to watch. Amazon recommends products you might want to buy. Twitter recommends users you might want to follow.

Methods to reccomend using data science

- Reccomending whats popular
- User based collaborative filtering
- Item-Based Collaborative Filtering

Other applications of data science

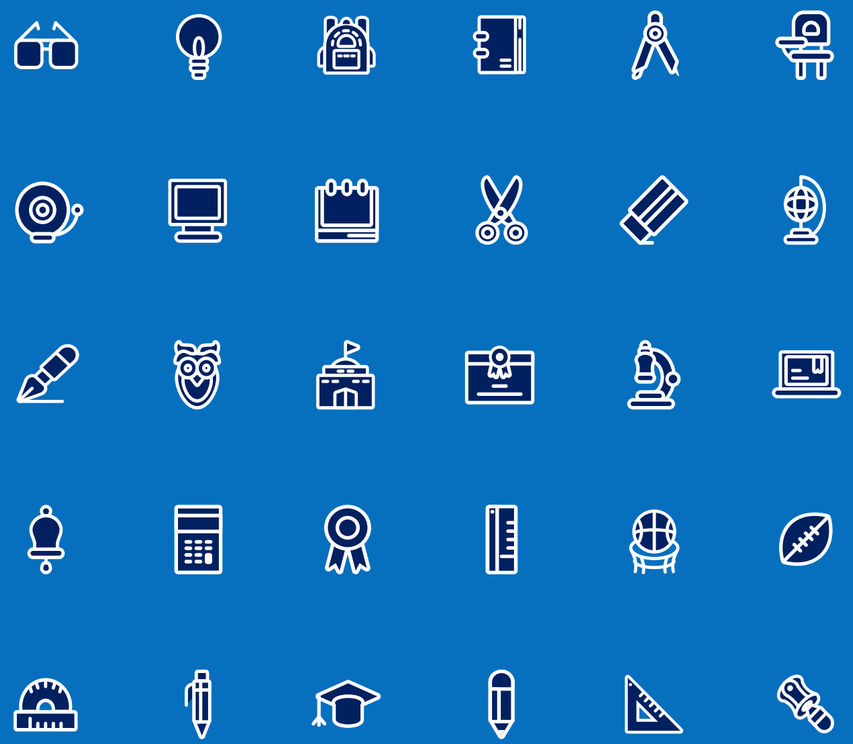
- Autocorrect
- Autocomplete
- Finance
- Virtual assistant
- OCR
- Chatbots



And now, once again, I bid my neophyte progeny go forth and prosper.

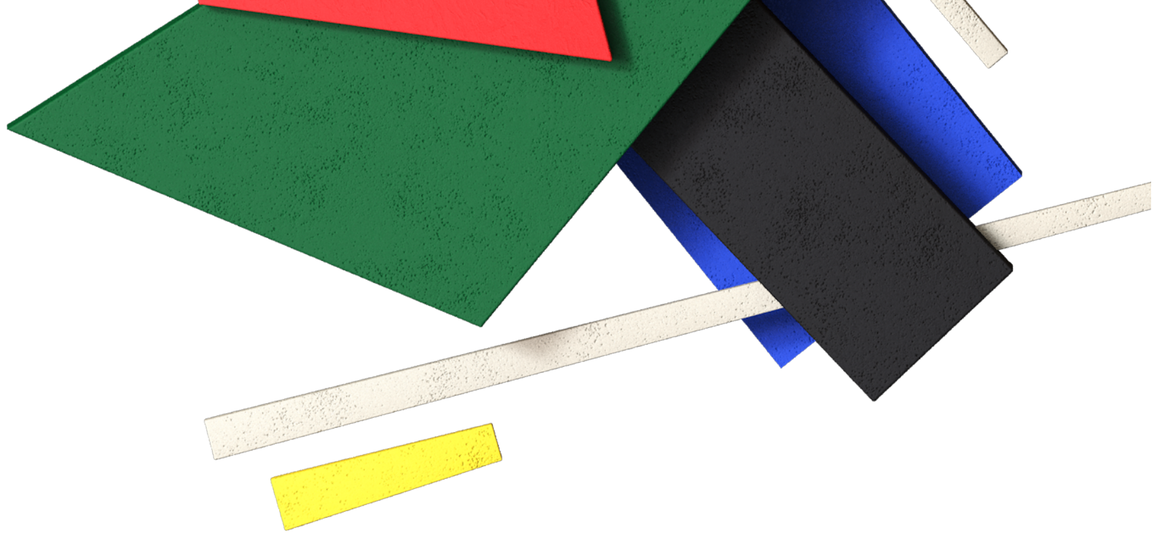
WHAT'S NEXT?

Where do you go from here? Assuming I haven't scared you off of data science, there are a number of things you should learn next.



What's next?

- Mathematics
- Not from Scratch
- NumPy
- pandas
- scikit-learn
- R
- Finding data
- Doing data science
- Looking at a Data Scientist's resume



Do you have any questions?

Ask us now or in the #general-help in the Discord server

To join our Discord community, visit the link given below:
https://coldorg.github.io/community/#discord