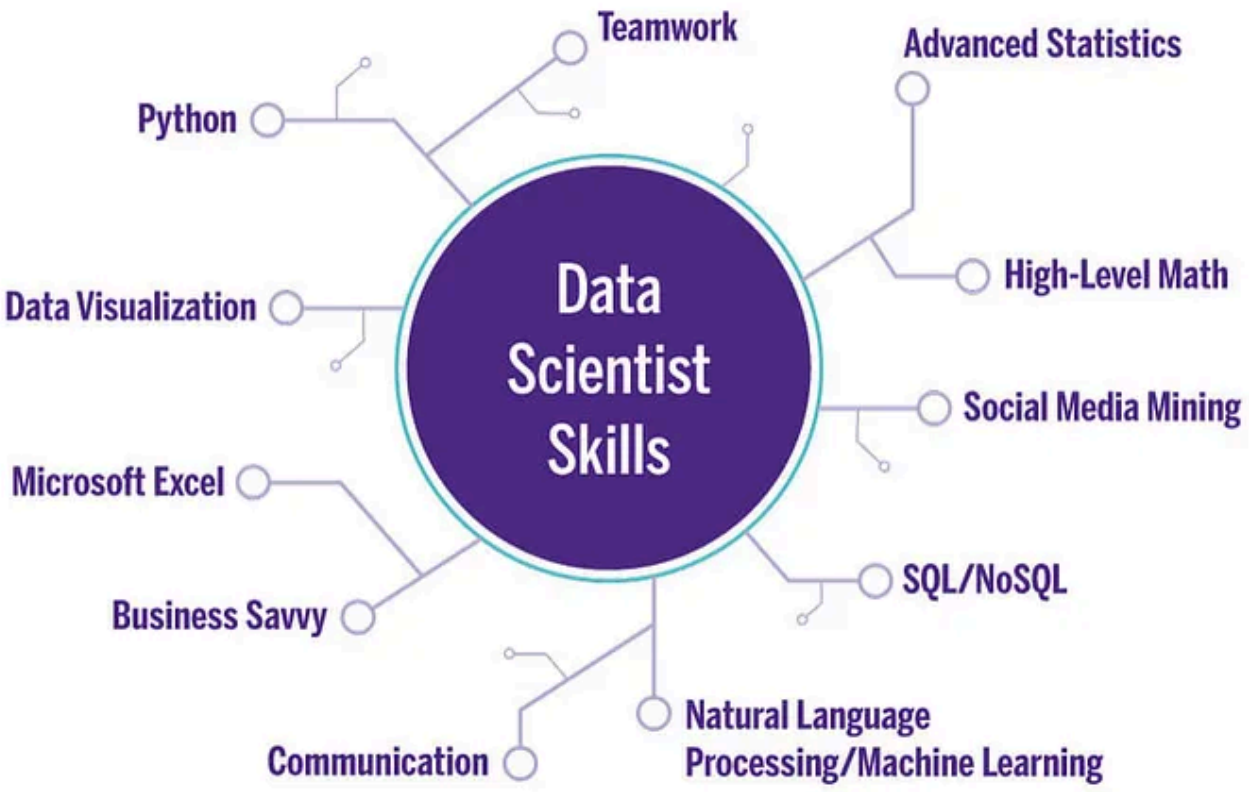
Master Data Science with This Comprehensive Cheat Sheet

Comprehensive Cheat Sheet for Data Science: Numpy, Pandas, Python, R, ML, DL, NLP, Stats, SQL, PySpark, Plotly, Seaborn, git, Excel, Tableau, and PowerBI

Ritesh Gupta · Follow

Published in Artificial Intelligence in Plain English · 6 min read · Jan 29, 2023

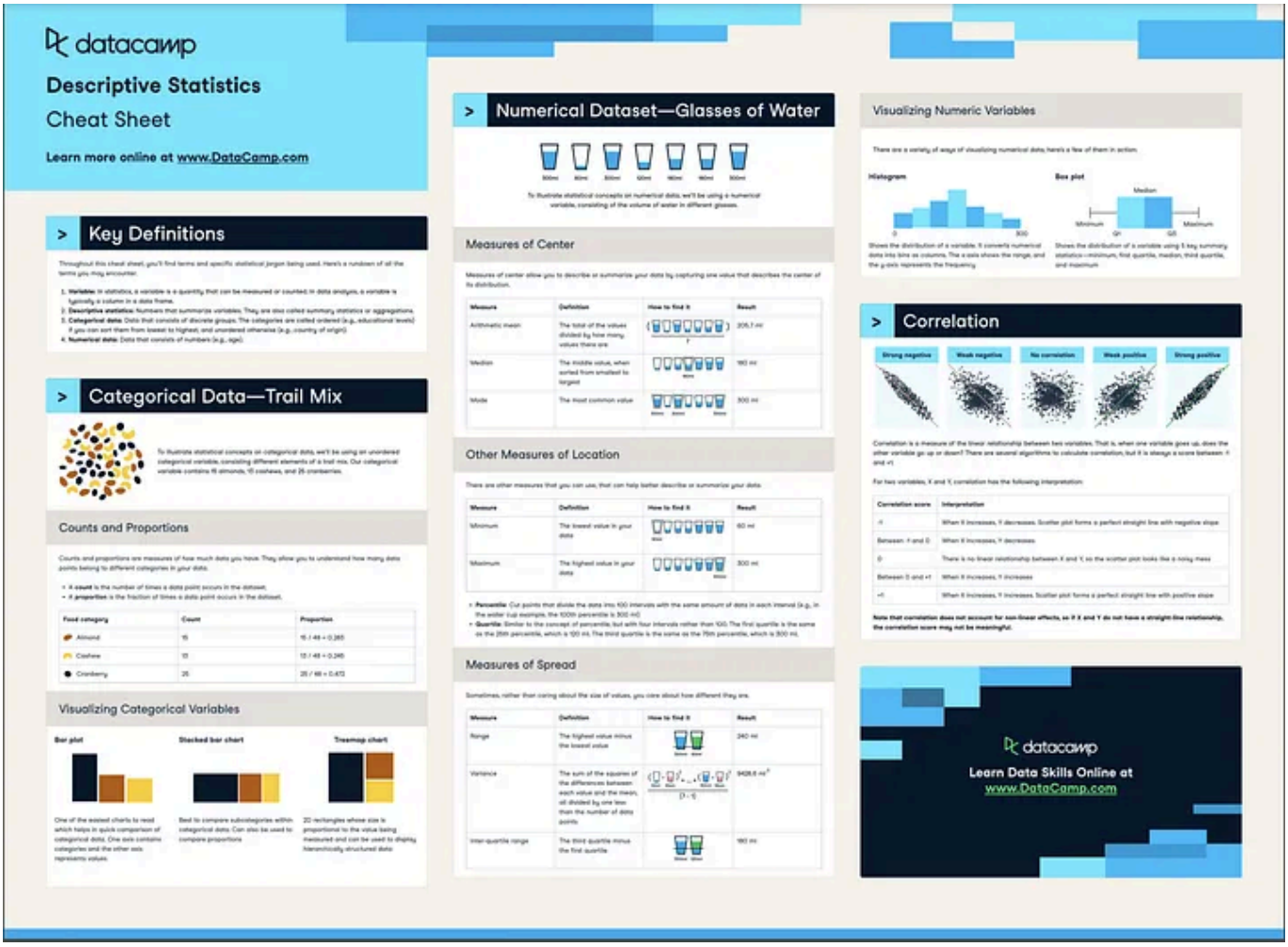
Data science is a rapidly growing field that combines statistics, mathematics, and computer science to extract insights and knowledge from data. As a data scientist, you need to be proficient in a variety of tools, techniques, and concepts to effectively analyze and visualize data. To help streamline your work, we have created the ultimate data science cheat sheet.



The cheat sheet covers all the essential topics in data science, from the basics of statistics and probability to advanced machine learning algorithms and deep learning techniques. It is designed to be a quick reference guide for data scientists, providing a comprehensive overview of the key concepts and tools used in the field.

Here are some of the topics covered in the data science cheat sheet:

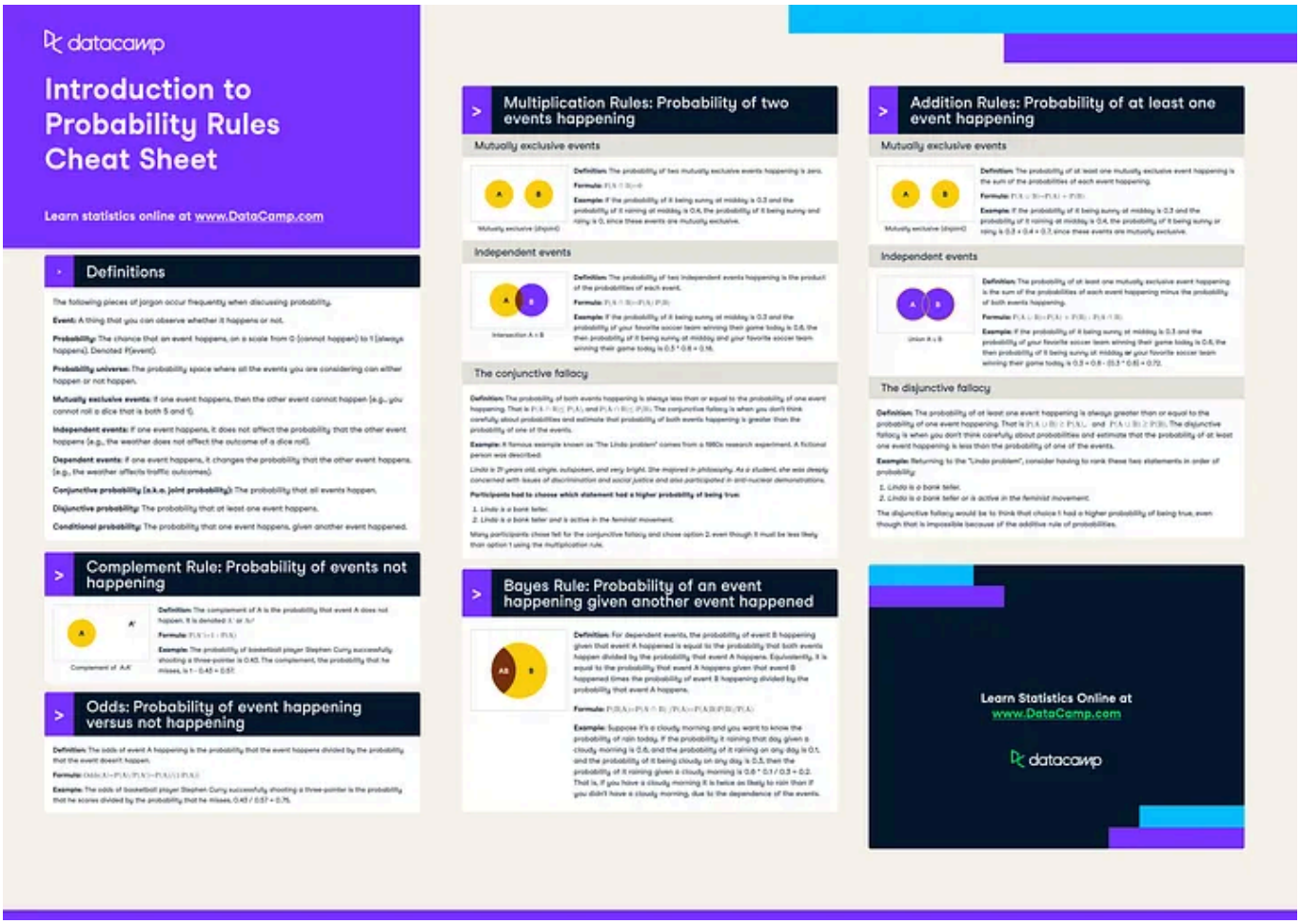
1. **Statistics**: Understanding the basics of statistics is crucial for data science. This section covers key concepts such as mean, median, mode, standard deviation, and correlation.



credit: datacamp

> *For Download Click here*

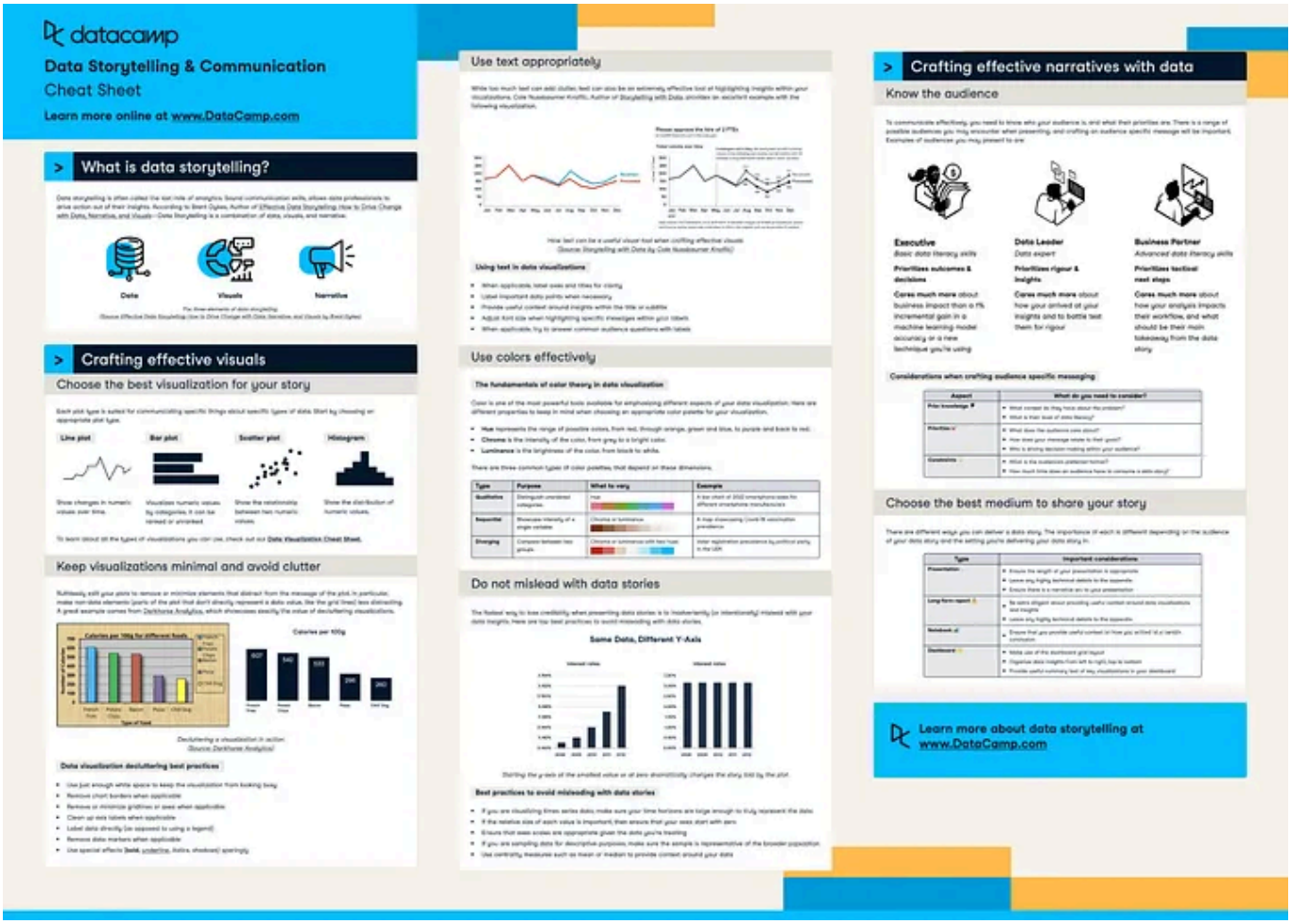
2. Probability: Probability is a fundamental concept in data science, used to make predictions and draw inferences from data. This section covers basic probability concepts, such as Bayes' theorem and conditional probability.



credit: datacamp

> *For Download Click here*

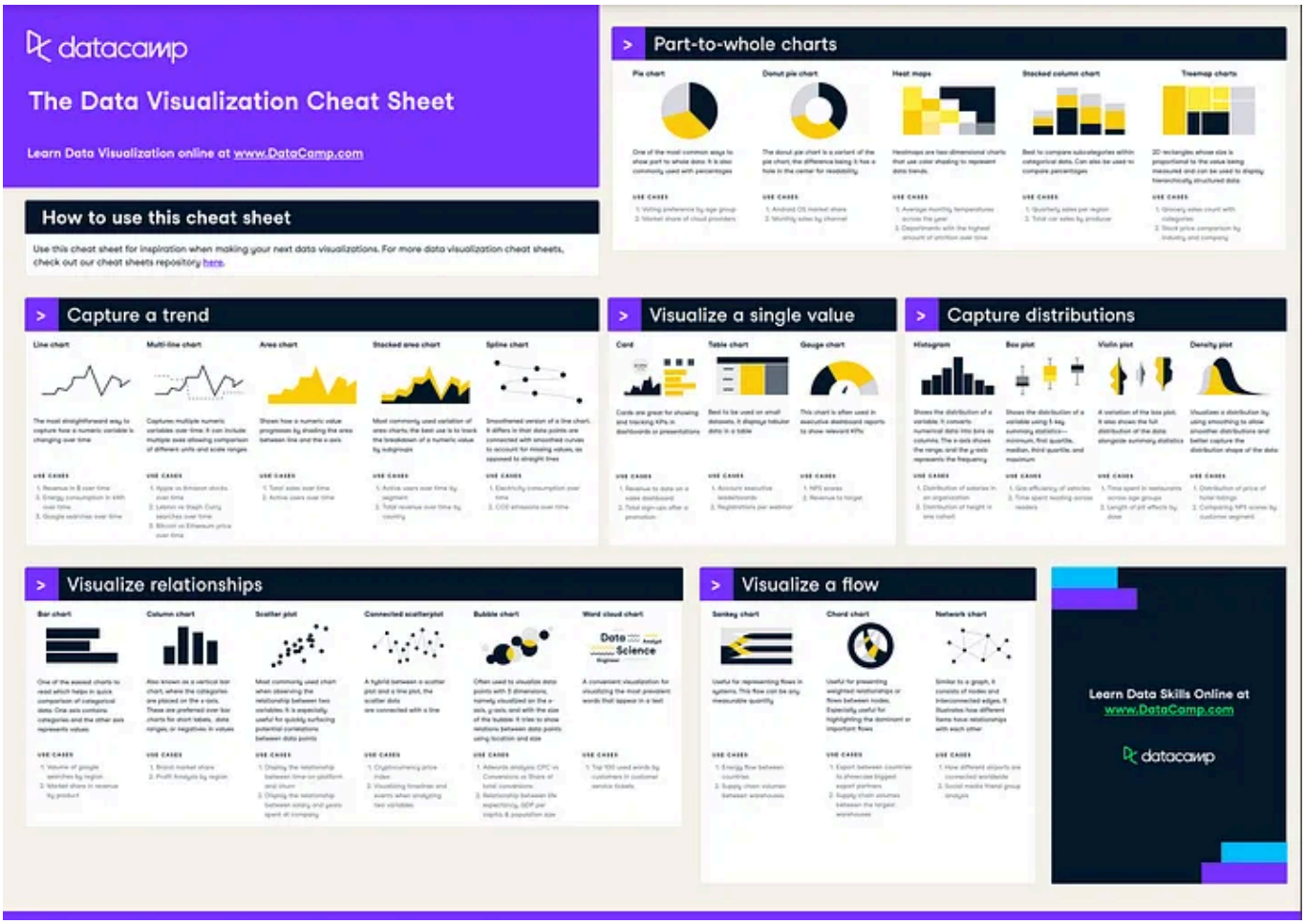
3. **Data storytelling:** Data storytelling is the process of communicating insights and findings from data analysis in a clear and compelling manner. The goal of data storytelling is to engage the audience and convey the key messages in a way that is easy to understand, memorable, and impactful.



credit: datacamp

> *For Download Click here*

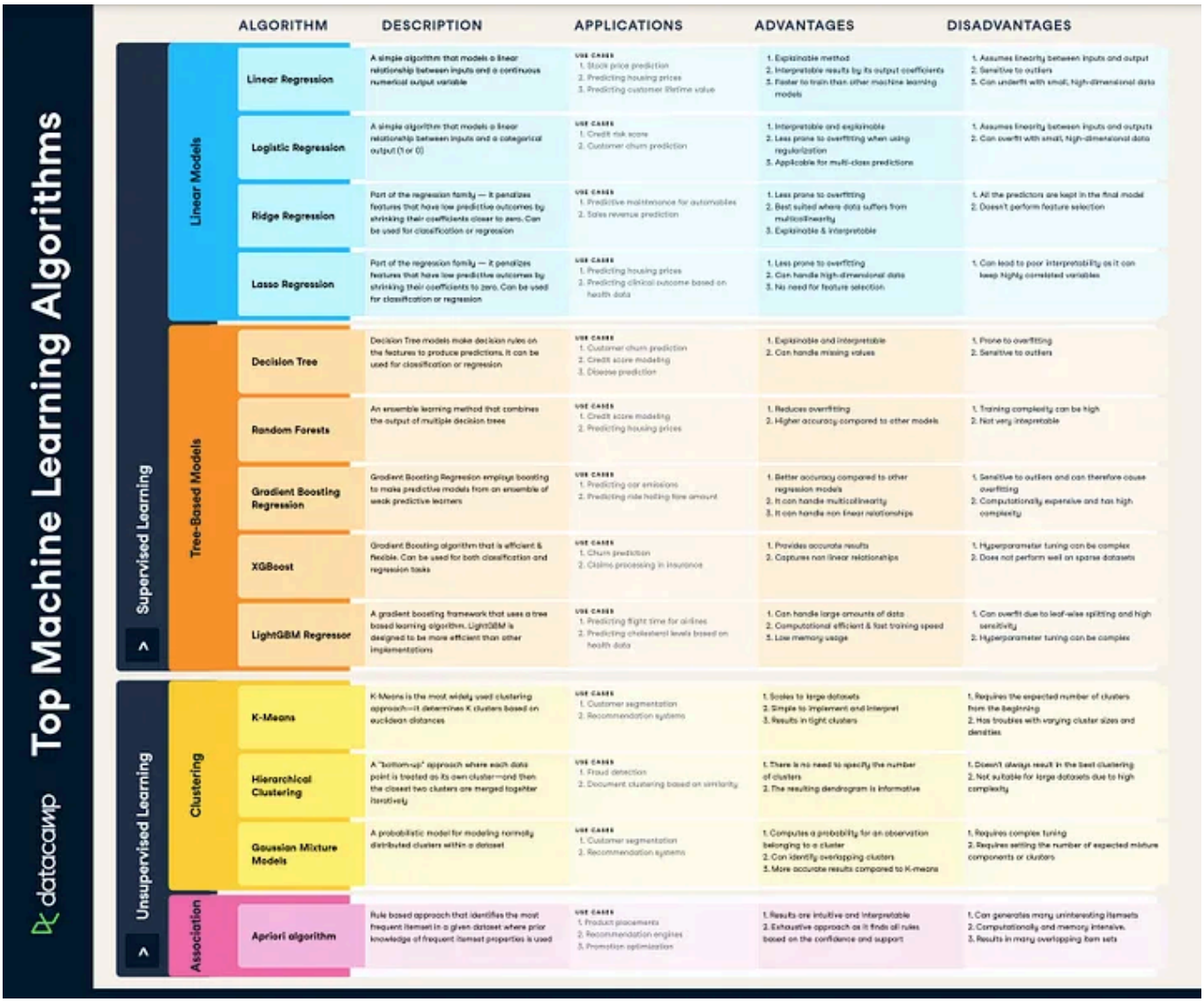
4. **Data Visualization:** Data visualization is an essential part of data science, allowing you to explore and understand the relationships and patterns in your data. This section covers popular visualization tools such as ggplot2, Matplotlib, and Seaborn.



credit: datacamp

> *For Download Click here*

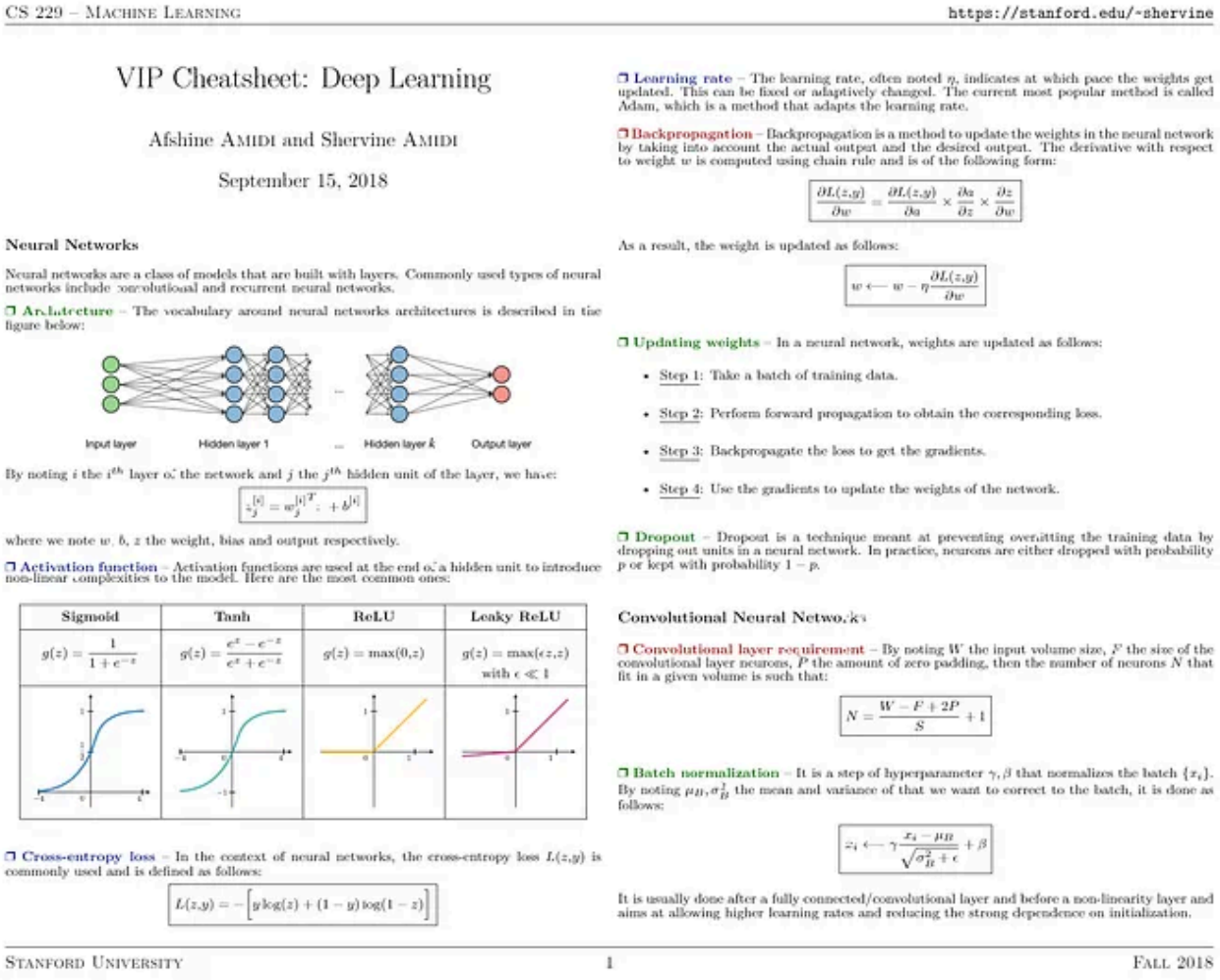
5. **Machine Learning:** Machine learning is the process of training algorithms to automatically learn from data and make predictions. This section covers popular machine learning algorithms such as linear regression, decision trees, and k-nearest neighbors.



credit: datacamp

> *For Download Click here*

6. **Deep Learning:** Deep learning is a subfield of machine learning that uses artificial neural networks to model complex relationships in data. This section covers popular deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).



CS 229 – Machine Learning https://stanford.edu/~shervine

VIP Cheatsheet: Deep Learning

Afshine Amidi and Shervine Amidi

September 15, 2018

Neural Networks

Neural networks are a class of models that are built with layers. Commonly used types of neural networks include convolutional and recurrent neural networks.

❒ **Architecture** – The vocabulary around neural networks architectures is described in the figure below:

Input layer Hidden layer 1 ... Hidden layer k Output layer

By noting i the i^{th} layer of the network and j the j^{th} hidden unit of the layer, we have:

$$z_j^{[i]} = w_j^{[i]T} \cdot + b^{[i]}$$

where we note w, b, z the weight, bias and output respectively.

❒ **Activation function** – Activation functions are used at the end of a hidden unit to introduce non-linear complexities to the model. Here are the most common ones:

Sigmoid	Tanh	ReLU	Leaky ReLU
$g(z) = \frac{1}{1+e^{-z}}$	$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$g(z) = \max(0,z)$	$g(z) = \max(\epsilon z, z)$ with $\epsilon \ll 1$

❒ **Cross-entropy loss** – In the context of neural networks, the cross-entropy loss $L(z,y)$ is commonly used and is defined as follows:

$$L(z,y) = -\Big[y \log(z) + (1-y)\log(1-z)\Big]$$

❒ **Learning rate** – The learning rate, often noted η, indicates at which pace the weights get updated. This can be fixed or adaptively changed. The current most popular method is called Adam, which is a method that adapts the learning rate.

❒ **Backpropagation** – Backpropagation is a method to update the weights in the neural network by taking into account the actual output and the desired output. The derivative with respect to weight w is computed using chain rule and is of the following form:

$$\frac{\partial L(z,y)}{\partial w} = \frac{\partial L(z,y)}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial w}$$

As a result, the weight is updated as follows:

$$w \longleftarrow w - \eta \frac{\partial L(z,y)}{\partial w}$$

❒ **Updating weights** – In a neural network, weights are updated as follows:

- Step 1: Take a batch of training data.
- Step 2: Perform forward propagation to obtain the corresponding loss.
- Step 3: Backpropagate the loss to get the gradients.
- Step 4: Use the gradients to update the weights of the network.

❒ **Dropout** – Dropout is a technique meant at preventing overfitting the training data by dropping out units in a neural network. In practice, neurons are either dropped with probability p or kept with probability $1-p$.

Convolutional Neural Networks

❒ **Convolutional layer requirement** – By noting W the input volume size, F the size of the convolutional layer neurons, P the amount of zero padding, then the number of neurons N that fit in a given volume is such that:

$$N = \frac{W - F + 2P}{S} + 1$$

❒ **Batch normalization** – It is a step of hyperparameter γ, β that normalizes the batch $\{x_i\}$. By noting μ_B, σ_B^2 the mean and variance of that we want to correct to the batch, it is done as follows:

$$x_i \longleftarrow \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

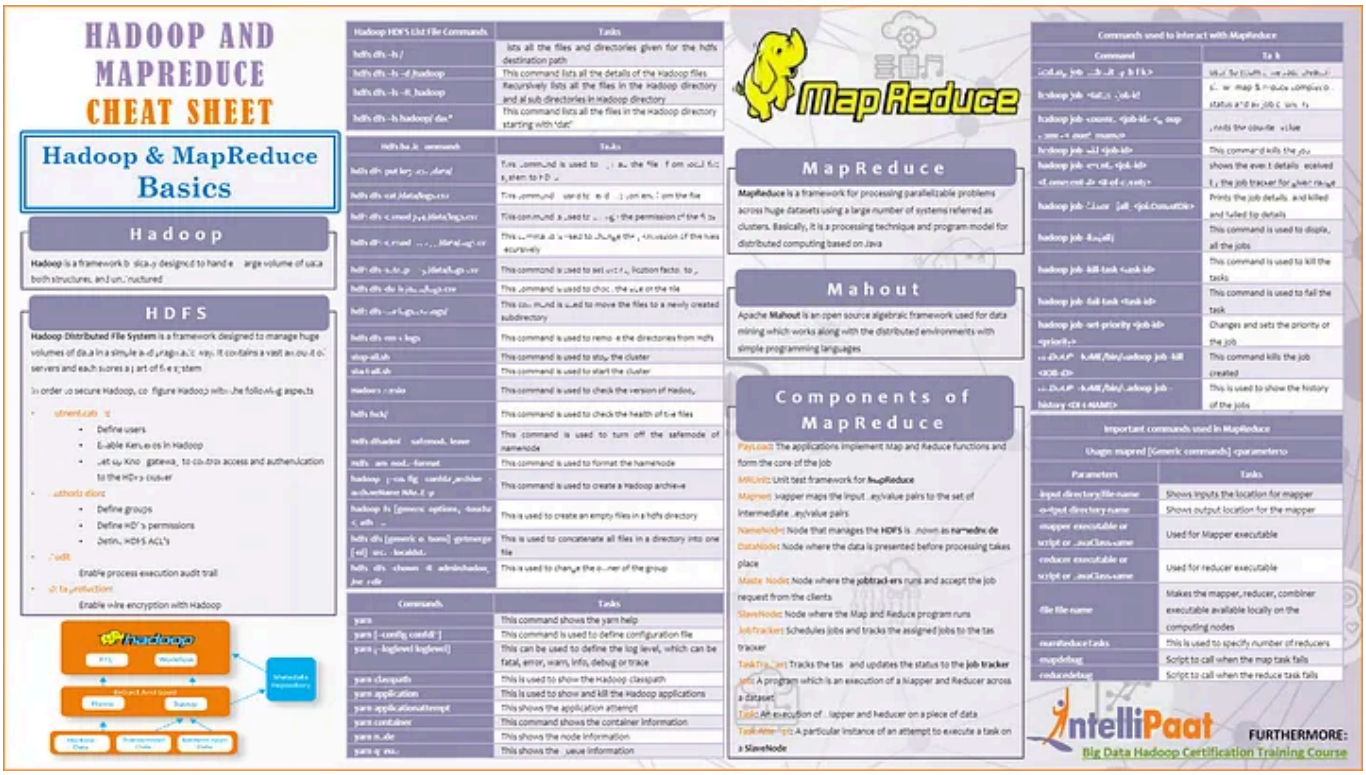
It is usually done after a fully connected/convolutional layer and before a non-linearity layer and aims at allowing higher learning rates and reducing the strong dependence on initialization.

Stanford University 1 Fall 2018

Credit: stanford

> *For Download Click here*

7. **Big Data:** Big data refers to large and complex data sets that can't be effectively processed by traditional data processing techniques. This section covers popular big data tools such as Hadoop, Spark, and NoSQL databases.



Credit: Intellipat

> *For Download Click here*

8. NLP: NLP (Natural Language Processing) is a subfield of artificial intelligence that focuses on the interaction between computers and humans using natural language. NLP involves the development of algorithms and models that enable computers to process, analyze, and understand human language. The goal of NLP is to make it possible for computers to understand and generate human language in a way that is both accurate and natural.

Cheatography

NLP Cheat Sheet

by sree017 via cheatography.com/126402/cs/24446/

Tokenization

Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. ... If the text is split into words using some separation technique it is called word tokenization and same separation done for sentences is called sentence tokenization.

```
# NLTK
import nltk
nltk.download('punkt')
paragraph = "write paragaraph here to convert into tokens."
sentences = nltk.sent_tokenize(paragraph)
words = nltk.word_tokenize(paragraph)
# Spacy
from spacy.lang.en import English
nlp = English()
sbd = nlp.create_pipe('sentencizer')
nlp.add_pipe(sbd)
doc = nlp(paragraph)
[sent for sent in doc.sents]
nlp = English()
doc = nlp(paragraph)
```

Tokenization (cont)

```
[word for word in doc]
# Keras
from keras.preprocessing.text import text_to_word_sequence
text_to_word_sequence(paragraph)
# genis
from gensim.summarization.textcleaner import split_sentences
split_sentences(paragraph)
from gensim.utils import tokenize
list(tokenize(paragraph))
```

Bag Of Words & TF-IDF

Bag of Words model is used to preprocess the text by converting it into a bag of words, which keeps a count of the total occurrences of most frequently used words

```
# counters = List of stences after pre processing like tokenization, stemming/lemmatization, stopwords
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 1500)
X = cv.fit_transform(counters).toarray()
```

Bag Of Words & TF-IDF (cont)

Term Frequency-Inverse Document Frequency (TF-IDF):

Term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

T.F = No of rep of words in setence/No of words in sentence

IDF = No of sentences / No of sentences containing words

```
from sklearn.feature_extraction.text import TfidfVectorizer
cv = TfidfVectorizer()
X = cv.fit_transform(counters).toarray()
```

N-gram Language Model:

An N-gram is a sequence of N tokens (or words). A 1-gram (or unigram) is a one-word sequence.the unigrams would simply be: "I", "love", "reading", "blogs", "about", "data", "science", "on", "Analytics", "Vidhya".

Bag Of Words & TF-IDF (cont)

A 2-gram (or bigram) is a two-word sequence of words, like "I love", "love reading", or "Analytics Vidhya". And a 3-gram (or trigram) is a three-word sequence of words like "I love reading", "about data science" or "on Analytics Vidhya".

Stemming & Lemmatization

From Stemming we will process of getting the root form of a word. We would create the stem words by removing the prefix of suffix of a word. So, stemming a word may not result in actual words.

```
paragraph = ""
# NLTK
from nltk.stem import PorterStemmer
from nltk import sent_tokenize
from nltk import word_tokenize
stem = PorterStemmer()
sentence = sent_tokenize(paragraph)[1]
words = word_tokenize(sentence)
[stem.stem(word) for word in words]
# Spacy
```

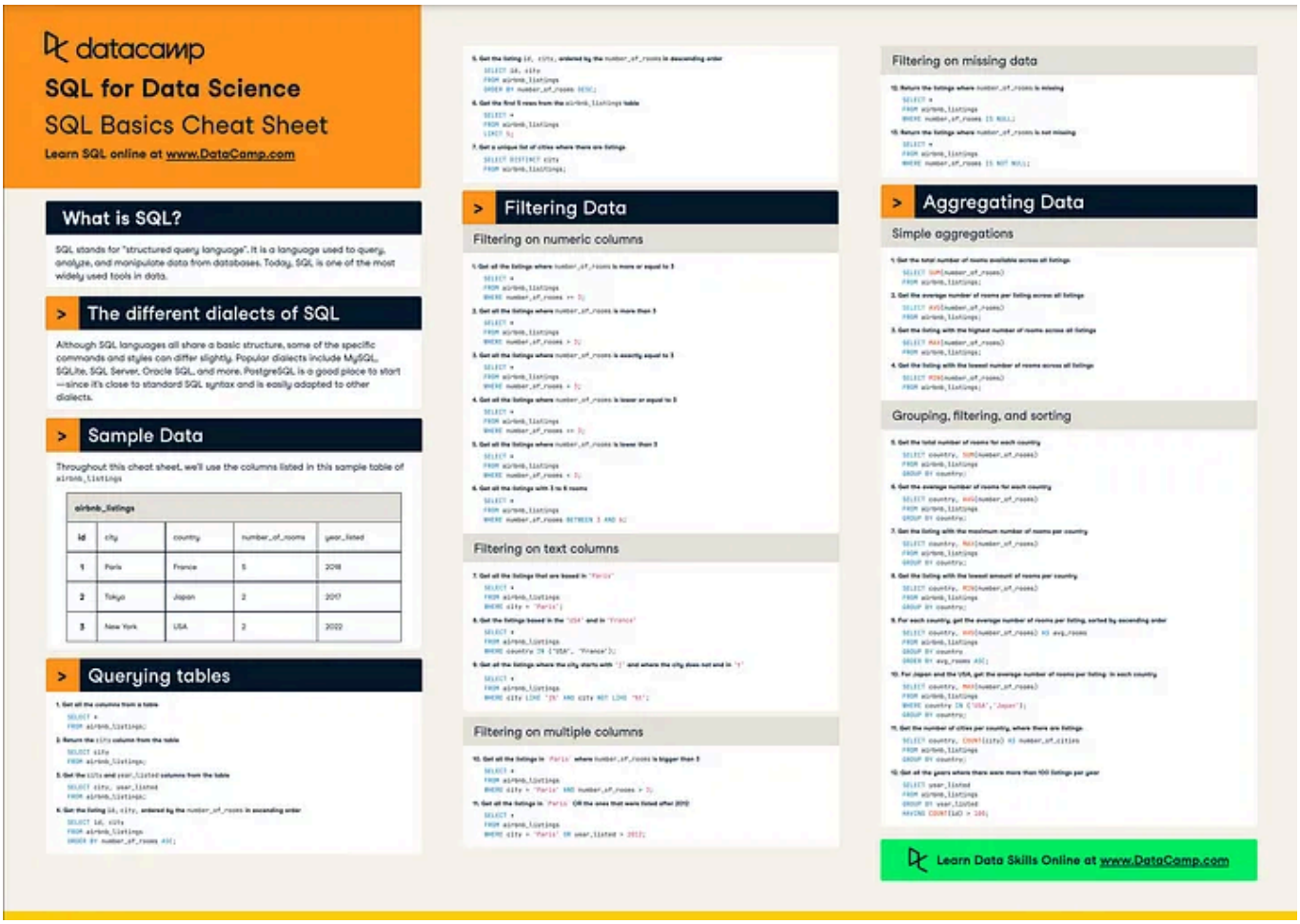
By sree017
cheatography.com/sree017/

Published 26th September, 2020.
Last updated 26th September, 2020.
Page 1 of 3.

credit: cheatography

> *For Download Click here*

9. **SQL:** SQL (Structured Query Language) is a programming language used to manage and manipulate data stored in relational databases. SQL is used to insert, update, and retrieve data in a database, as well as to create and modify database structures, such as tables and indexes.



Credit: datacamp

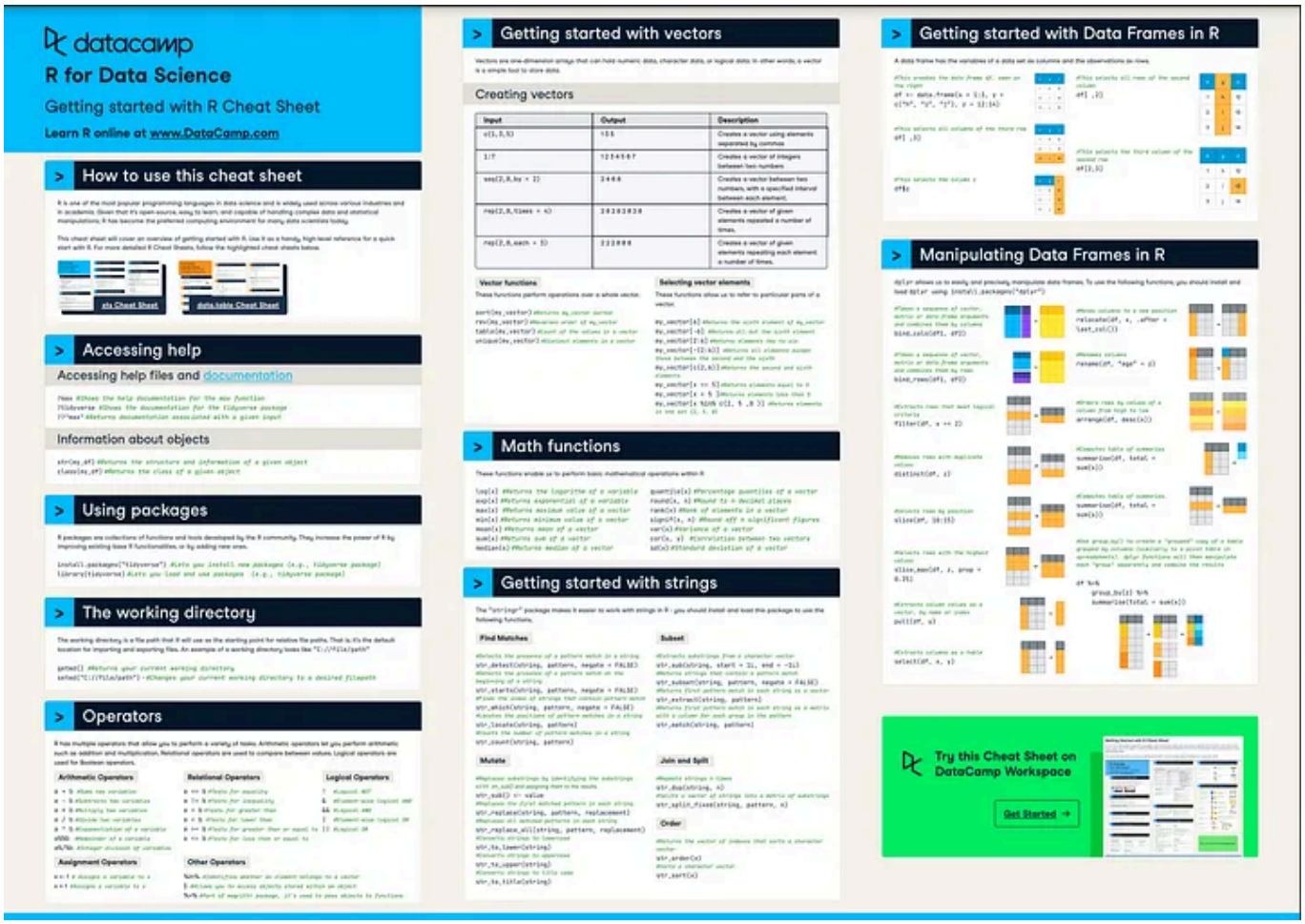
> *For Download Click here*

10. **Python:** Python is a high-level programming language that is widely used for a variety of tasks, including web development, data analysis, artificial intelligence, and scientific computing. Python is known for its simple and expressive syntax, making it a popular choice for both beginners and experienced developers.

credit: datacamp

> *For Download Click here*

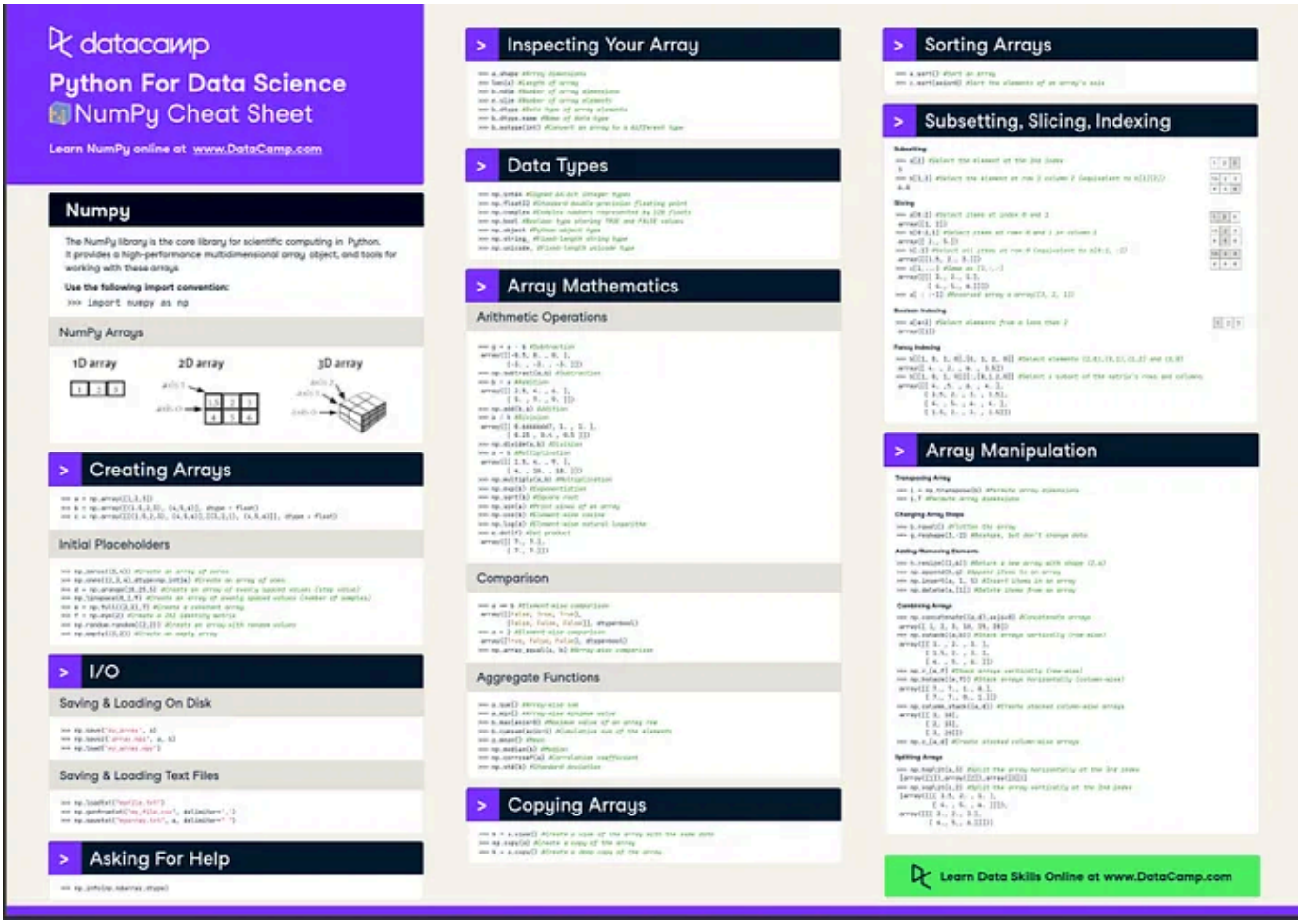
11. R: R is a high-level programming language and software environment for statistical computing and graphics. R is widely used by statisticians, data scientists, and researchers for data analysis and visualization.



credit: datacamp

> *For Download Click here*

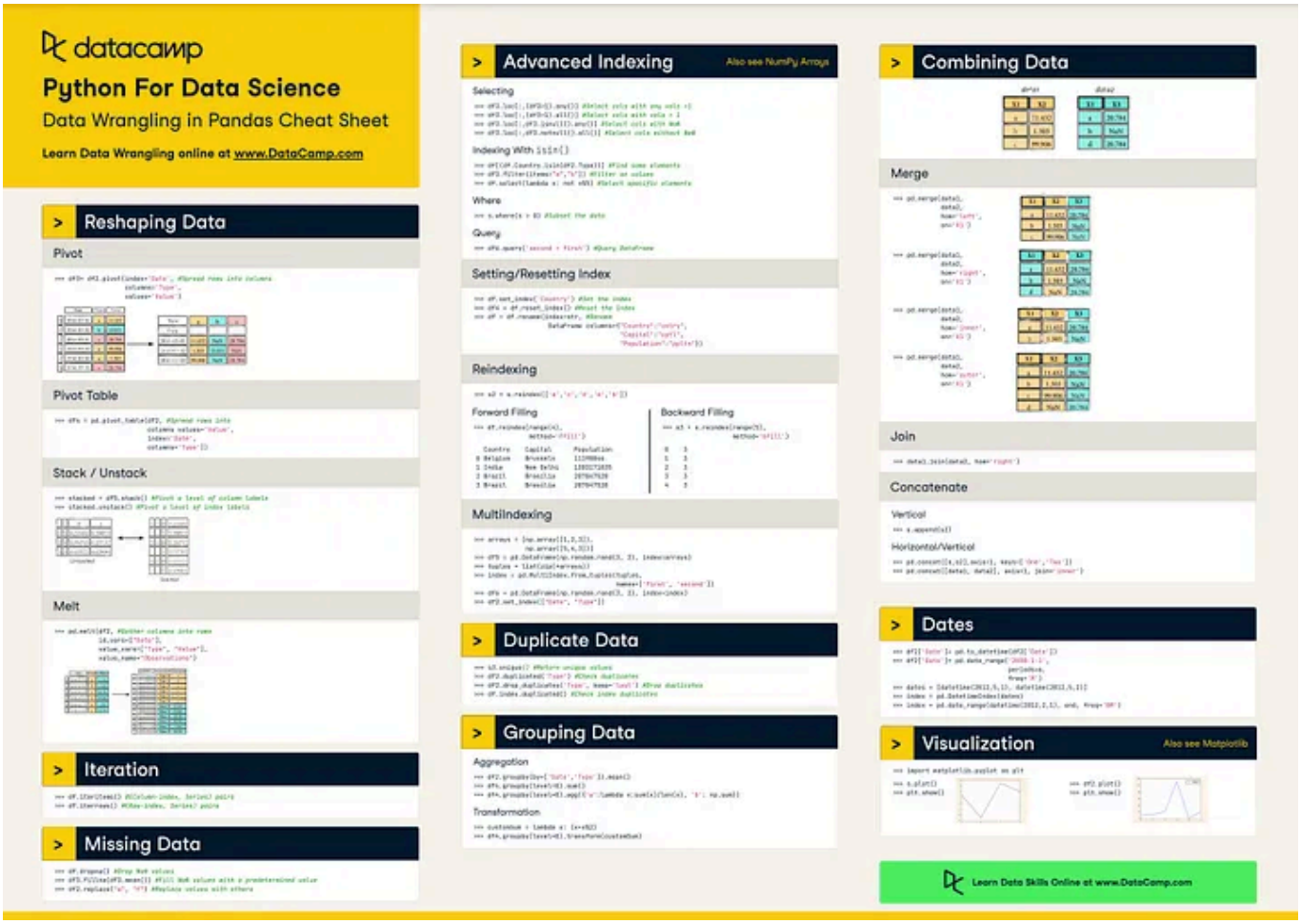
12. Numpy: NumPy is a Python library for numerical computing, specifically for arrays and matrices. It is a fundamental package for scientific computing with Python, and is widely used in data science, machine learning, and other technical fields.



credit: datacamp

> *For Download Click here*

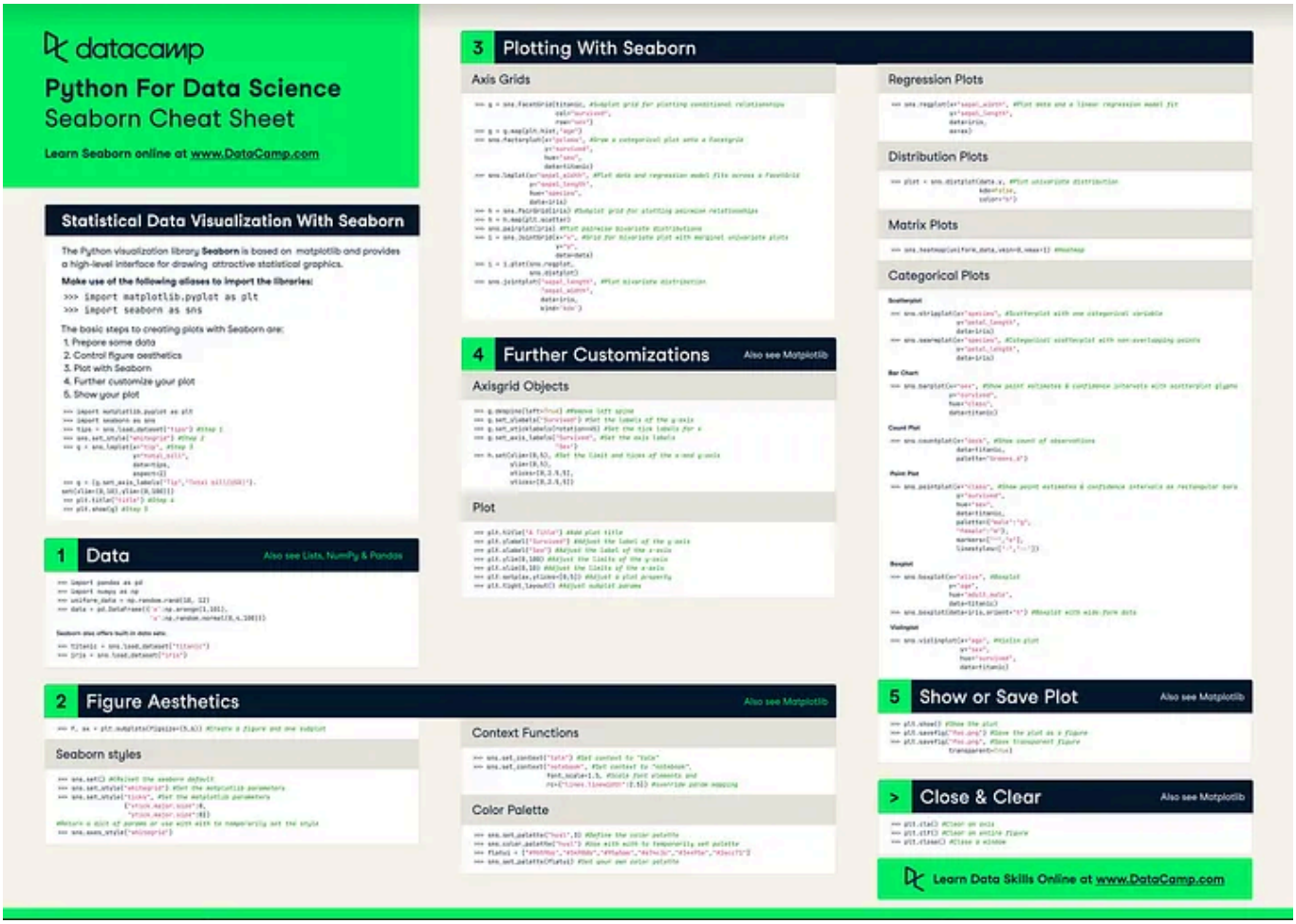
13. Pandas: Pandas is a Python library for data manipulation and analysis. It provides data structures for efficiently storing large datasets and tools for working with them. Pandas is widely used in data science, machine learning, and other technical fields for tasks such as data cleaning, aggregation, and transformation.



Credit: datacamp

> *For Download Click here*

14. Seaborn: Seaborn is a data visualization library in Python, built on top of Matplotlib, that provides a high-level interface for creating statistical graphics. It is focused on the use of visualizations for exploring and understanding the structure of complex datasets.



credit:datacamp

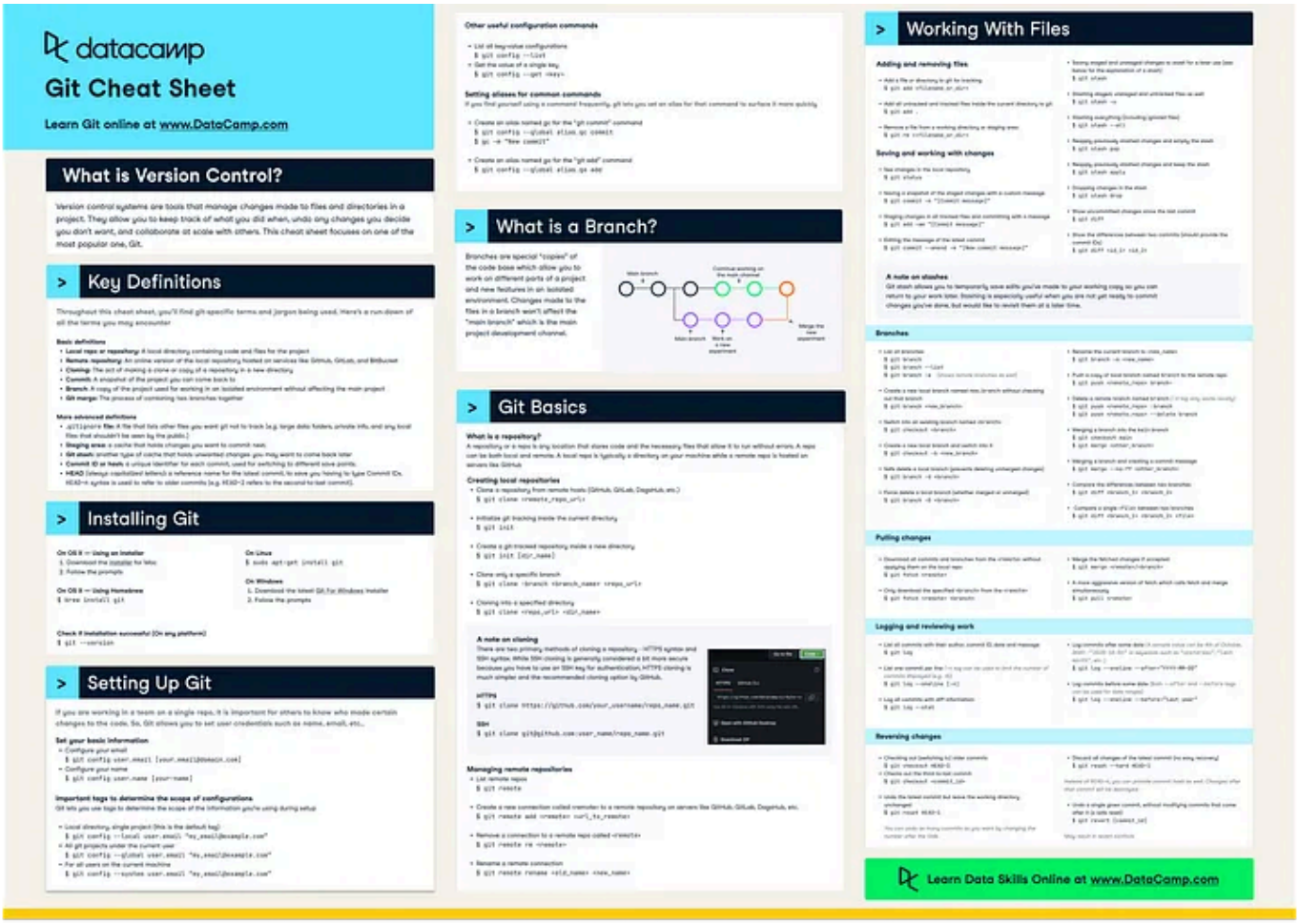
> *For Download Click here*

15. Plotly Express: Plotly Express is a high-level data visualization library in Python, built on top of Plotly, that provides a simple and expressive way to create interactive, animated and publication-quality visualizations. It is designed to help users quickly create visualizations without writing too much code and focuses on providing a wide range of charts and options with sensible defaults.

credit:datacamp

> *For Download Click here*

16. Git: Git is a version control system for software development and code management. It allows developers to track changes made to code over time, collaborate on projects with other developers, and maintain different versions of the codebase. Git operates on a distributed model, meaning that multiple copies of a repository can exist on different machines, making it easy to work offline and share changes with others.



credit: datacamp

> *For Download Click here*

17. PySpark: PySpark is a Python API for Apache Spark, an open-source, distributed computing system for big data processing and analysis. PySpark provides a way for Python developers to use Spark's powerful processing engine to process and analyze large datasets in parallel. It enables developers to scale out their computations and perform complex data transformations and aggregations, while leveraging the simplicity and expressiveness of Python programming.

credit: datacamp

> *For Download Click here*
>
> 18. ***Excel*** *Cheat Sheet* ***For Download Click here***
>
> 19. ***Tableau*** *Cheat Sheet* ***For Download Click here***
>
> 20. ***Power BI*** *Cheat Sheet* ***For Download Click here***

Conclusion

In conclusion, the data science cheat sheet is a valuable resource for anyone looking to expand their knowledge in the field of data science. Whether you're a beginner or an experienced data scientist, the cheat sheet provides a

quick reference for all the essential concepts and tools used in the field. Bookmark this cheat sheet and keep it handy as you work on your next data science project.

Thanks for Reading!

If you enjoyed this, follow me to never miss another article on data science guides, tricks and tips, life lessons, and more!

*More content at **PlainEnglish.io**. Sign up for our **free weekly newsletter**. Follow us on **Twitter**, **LinkedIn**, **YouTube**, and **Discord**.*

*Interested in scaling your software startup? Check out **Circuit**.*

Data Visualization | Data Science | Machine Learning | Deep Learning | Python

Written by Ritesh Gupta

3.4K Followers · Writer for Artificial Intelligence in Plain English

Data Scientist, I write Article on Machine Learning| Deep Learning| NLP | Open CV | AI Lover ❤️

Follow

More from Ritesh Gupta and Artificial Intelligence in Plain English

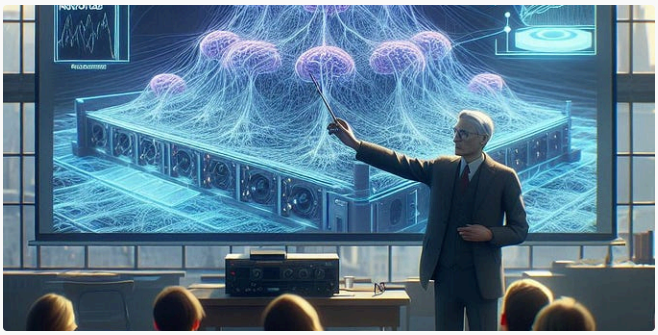
Ritesh Gupta

10 Automated EDA Tools That Will Save You Hours Of Work

Exploratory Data Analysis (EDA) is the process of analyzing and summarizing the...

7 min read · Jan 29, 2023

386 6



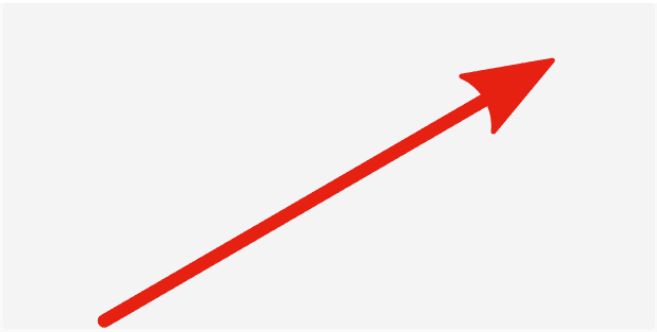
Austin Star... in Artificial Intelligence in Plain Engli...

Reinforcement Learning is Dead. Long Live the Transformer!

Large Language Models are more powerful than you imagine

· 8 min read · Jan 13, 2024

2.1K 49



Kane Hoop... in Artificial Intelligence in Plain Engli...

Understanding AI Similarity Search

A guide to understanding similarity search (also known as semantic search), one of the...

12 min read · Apr 16, 2024

169 3

Ritesh Gupta

14 Life Changing Lessons From Chanakya Niti everyone should...

Who is Chanakya?

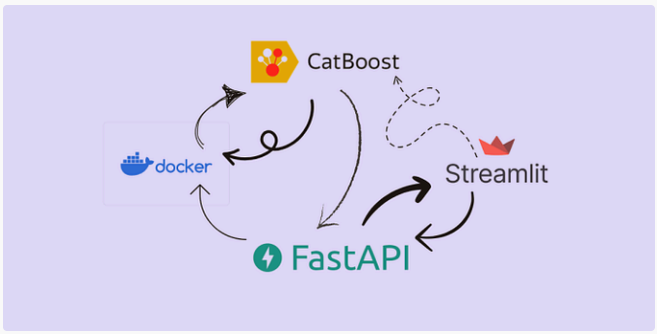
4 min read · Jan 29, 2023

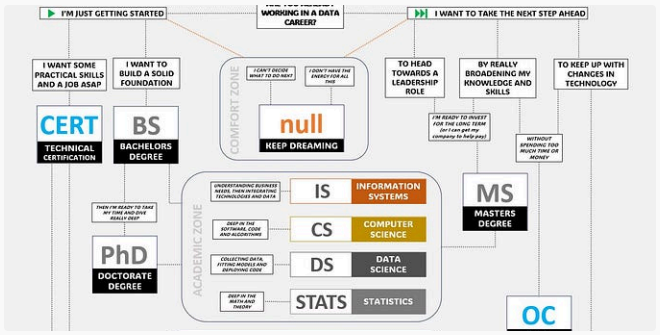
108 1

See all from Ritesh Gupta

See all from Artificial Intelligence in Plain English

Recommended from Medium





Ramazan Olmez

End-to-End Machine Learning Project: Churn Prediction

The main objective of this article is to develop an end-to-end machine learning project. For...

18 min read · Feb 22, 2024

131

Stan Pugsley

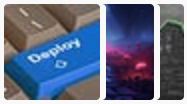
Roadmap for Your Data Career

As you design your career in data, you've got to avoid getting stuck in your comfort zone o...

5 min read · Jan 31, 2024

90

Lists



Predictive Modeling w/ Python

20 stories · 1134 saves

Practical Guides to Machine Learning

10 stories · 1360 saves

Coding & Development

11 stories · 581 saves

Natural Language Processing

1410 stories · 906 saves

Nathan Rosidi

Data Science in 2024—What Has Changed

What has changed in the data science landscape, and what are the challenges of th...

4 min read · Jan 29, 2024

1.3K 24

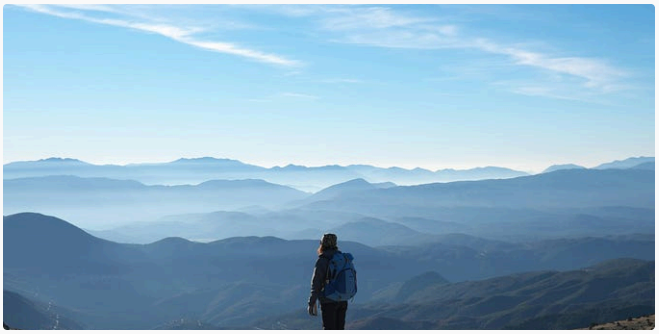
Nilimesh Halder, PhD in Level Up Coding

Unlocking Economic Insights with Exploratory Data Analysis

A Comprehensive Guide with Python and R

· 24 min read · 6 days ago

25



Sze Zhong LIM in Data And Beyond

Mastering Exploratory Data Analysis (EDA): Everything You...

A systematic approach to EDA your data and prep it for machine learning.

18 min read · Apr 6, 2024

453 4

Miriam Santos in Towards Data Science

A Data Scientist's Essential Guide to Exploratory Data Analysis

Best Practices, Techniques, and Tools to Fully Understand Your Data

11 min read · May 30, 2023

1.3K 13