DS/AI Self-Starter Handbook

BUILD YOUR OWN ROADMAP

Ankit Rathi

From a time around when DS/AI field started picking up, every other day I get at least 8–10 messages from DS/AI starters & enthusiasts on ‘How can I get into DS/AI field?’. Over a while, I have improvised my response based on the follow-up questions they ask like:

1. What is the difference between DS, ML, DL, AI, DM?

2. What are the roles in DS/AI, who does what?

3. What concepts, processes & tools they need to learn?

4. Which books, courses, etc they need to refer to?

5. How to build a DS/AI portfolio?

6. How to write a resume for DS/AI?

7. How to build a helpful network?

8. How to search for the job?

9. How to prepare for the interview?

10. How to stay up to date in this still-evolving field?

You can notice that these questions are not conceptual ones and there is no dedicated material to address these roadblocks. I thought why not to build a framework or a road-map for DS/AI starters and enthusiasts so that I need not to answer the same type of questions again and again. And that is when I started documenting what a starter or enthusiast need to do step by step in order to reach a level when he is ready to tackle any challenge thrown to him. My answer to the above questions in a structured way to help DS/AI starters & enthusiasts is this book. This book covers the framework to launch your DS/AI career in 8 chapters.

Ankit Rathi provides unique combination of Data Engineering (DB/ETL/DWH/BI)/Architecture (Data Management & Governance) & Data Science (ML/DL/AI) with more than a decade of demonstrated history of working in IT industry using Data & Analytics. His interest lies primarily in building end to end DS/AI applications/products following best practices of Data Engineering and Architecture.

In his free time, he blogs about various topics on DS/AI field & tries to simplify it for starters & enthusiasts.

ankitrathi.com

Rs 499/-
ISBN 978-1-64587-432-4

DS/AI Self-Starter Handbook

Build Your Own Roadmap

Ankit Rathi

ankitrathi.com

To my wife, Divya, who's always accepted me the way I am and supported my hustle, drive & ambition.

To my children, Aarsh & Driti, who are the reason to wake up every morning and work as hard as I can.

DS/AI Self-Starter handbook is a great resource for aspirants starting in the space of Data Science. It covers approach and useful resources that can help in your learning journey and written by one who himself is an Data Science practitioner. I recommend this to anyone who are aspiring to get into Data Science and are looking for insights on how and where to get started.

Srivatsan Srinivasan
Chief Data Scientist (Cognizant)

Wow, this is very impressive! It has taken some time to review, but WOW!
I should have had you as a co-author next time!!!

T. Scott Clendaniel
Chief Data Scientist (Legg Mason)

To be great data scientist you should emphasis on skillset and mindset. Where a lot of book that give you skill set, this is the first book I read that dedicating to shape data scientist mindset.

Nabih Ibrahim Bawazir
Data Science Head (Datanest)

Extremely laudable & heroic attempt to put all your thoughts and experience together to help people.

Sumit Pal
Big Data Architect (Qcentive)

Ankit has done a great job summarizing what is possibly one of the toughest and most frequently asked questions, "How to get started with data science?". Packed with information, this book will definitely be helpful for people from both academia and industry looking to get started on their own Data Science and AI journey.

Dipanjan Sarkar
Data Scientist (Rad Hat)

I think it is a brilliant book for starting Career in Data Science as New Entrants to Data Science often deviate from Path to reach End Goal and this Book tries to solve that Problem in a easy way. I would really like to Congratulate Ankit for Providing Data Science Career Steps in this useful manner.

Yatin Bhatia
Data Scientist (RxLogix)

An indispensable guide and a valuable resource for anyone seeking to enter the field of Data Science. Replete with great advice directly from the author's personal experience.

Parul Pandey
Data Science Evangelist (H2O.ai)

This book kicks you into the right direction definitely worth reading for the beginners trying to break into DS/AI.

Avik Jain
Machine Learning Intern (EMA Solutions)

If you are one among people struggling to identify the right book for data science, this book would probably help to understand where to start, how to prepare, how to develop the habit of continuous learning.

Vishnu Durgha Prasaad
Data Science Practitioner

About the Author

Ankit Rathi is currently working as a Lead Architect-DS/AI at SITA aero. He is a Data Science (ML/DL/AI) practitioner with more than a decade of demonstrated history of working in IT industry using Data & Analytics. His interest lies primarily in the theory & application of artificial intelligence, particularly in developing business applications for machine learning and deep learning. Ankit's work at SITA aero has revolved around designing FlightPredictor product & building the CoE capability. During his tenure as a Principal Consultant at Genpact HCM, Ankit architected and deployed machine learning pipelines for various clients across different industries like Insurance, F&A. He was previously a Tech Lead at RBS IDC where he designed and developed various data intensive applications in AML & Mortgages area. Ankit is a well-known author for various publications (Towards Data Science, Analytics Vidhya etc) on Medium where he actively contributes by writing blog-posts on concepts & latest trends in Data Science. His blog-series on 'Probability & Statistics for Data Science' has been well received by Data Science community in 2018. He is followed by around 30K data science practitioners & enthusiasts on LinkedIn.

U0.1: Webpage: *https://www.ankitrathi.com/*

Table of Contents

Build

DS/AI: Self-Starter Kit

Build Your Own Roadmap

Chapter 3

Working on Building Blocks

There are few core skills of every job. To perform that job, you need to be aware of core concepts, you need to be aware of the end to end process and you need to learn how to use related tools to perform that job. Data science in no different job, it has its own core concepts, processes and tools.

This chapter covers the core concepts you need to learn, end-to-end process you need to be aware of & important tools you need to master to work as a data scientist.

Concepts
Process
Tools

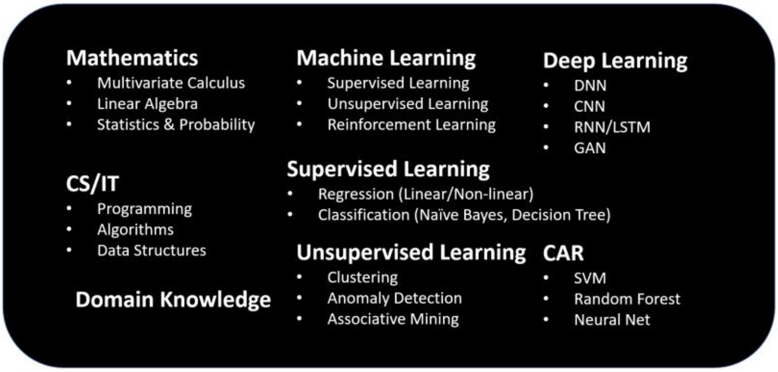
Working on the Building Blocks
DS/AI: Self-Starter Kit

Please note that this chapter only outlines the concepts, processes and tools used by data scientists. I have covered the resources (mostly free) for these topics in upcoming chapter.

Still, if you want to build a quick understanding you can refer following link:

U03: DS/AI The Complete Reference: *https://medium.com/data-deft/data-science-the-complete-reference-series-3fb35077fc5a*

3.1 Concepts to Learn



Mathematics
- Multivariate Calculus
- Linear Algebra
- Statistics & Probability

Machine Learning
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Deep Learning
- DNN
- CNN
- RNN/LSTM
- GAN

CS/IT
- Programming
- Algorithms
- Data Structures

Supervised Learning
- Regression (Linear/Non-linear)
- Classification (Naïve Bayes, Decision Tree)

Domain Knowledge

Unsupervised Learning
- Clustering
- Anomaly Detection
- Associative Mining

CAR
- SVM
- Random Forest
- Neural Net

Concepts to Learn

Working on Building Blocks

Mathematics

Data science contains math — no avoiding that! This section is for learners about basic math they need in order to be successful in almost any data science project/problem. So let's start:

Multivariate Calculus

Calculus is a set of tools for analyzing the relationship between functions and their inputs. In Multivariate Calculus, we can take a function with multiple inputs and determine the influence of each of them separately.

In data science, we try to find the inputs which enable a function to best match the data. The slope or descent describes the rate of change off the output with respect to an input. Determining the influence of each input on the output is also one of the critical tasks. All this requires a solid understanding of Multivariate Calculus.

Linear Algebra

The word *algebra* comes from the Arabi word "*al-jabr*" which means "*the reunion of broken parts*". This is the collection of methods deriving unknowns from knowns in mathematics. *Linear Algebra* is the branch that deals with *linear equations* and *linear functions* which are represented through *matrices* and *vectors*. In simpler words, it helps us understand geometric terms such as planes, in higher dimensions, and perform mathematical operations on them. By definition, algebra deals primarily with scalars (one-dimensional entities), but Linear Algebra has vectors and matrices (entities which possess two or more dimensional components) to deal with linear equations and functions.

Linear Algebra is central to almost all areas of mathematics like *geometry* and *functional analysis*. Its concepts are a crucial prerequisite for understanding the theory behind *Data Science*. You don't need to understand *Linear Algebra* before getting started in *Data Science*, but at some point, you may want to gain a better understanding of how the *different algorithms* really work under the hood. So if you really want to be a professional in this field, you will have to master the parts of *Linear Algebra* that are important for *Data Science*.

Statistics & Probability

Statistics is a mathematical body of science that pertains to the *collection*, *analysis*, *interpretation* or *explanation*, and *presentation* of data. Probability is the chance that something will happen — how likely it is that some event will happen.

Statistics help you to understand your data and is an initial & very important step of Data Science. This is due to the fact that Data Science is all about making predictions and you can't predict if you can't understand the patterns in existing data.

Uncertainty and randomness occur in many aspects of our daily life and having a good knowledge of probability help us make sense of these uncertainties. Learning about probability helps us make informed judgments on what is likely to happen, based on a pattern of data collected previously or an estimate.

Data science often uses statistical inferences to predict or analyze trends from data, while statistical inferences use probability distributions of data. Hence knowing probability & statistics and its applications are important to work effectively on data science problems.

Programming

To execute the DS/AI pipeline, you need to learn algorithm design as well as fundamental programming concepts such as data selection, iteration and functional decomposition, data abstraction and organisation. In addition to this, you need to learn how to perform simple data visualizations using programming and embed your learning using problem-based assignments.

Machine Learning Algorithms

Machine learning algorithms can be divided into 3 broad categories —

- Supervised learning,
- Unsupervised learning
- Reinforcement learning

Supervised learning is useful in cases where a property (*label*) is available for a certain dataset (*training set*) but is missing and needs to be predicted for other instances. Unsupervised learning is useful in cases where the challenge is to discover implicit relationships in a given *unlabelled* dataset (items are not pre-assigned). Reinforcement learning falls between these 2 extremes — there is some form of feedback available for each predictive step or action, but no precise label or error message.

Intrinsic details of various algorithms is not in scope of this book, you can refer the resources mentioned in the next chapter to learn them.

Supervised learning can be further divided into Regression (Linear, Non-linear etc) & Classification (Logistics Regression, Decision Tree, Naïve Bayes etc) algorithms. Some algorithms can be used for regression as well as classification i.e. Random Forests, Support Vector Machines etc.

Unsupervised learning can also be further divided into Clustering, Anomaly Detection, Associative Mining.

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

Deep Learning Frameworks

Deep learning frameworks are a more advanced form of ML and solve specific problems where data is either unstructured or huge or both. Neural Nets, CNNs, RNNs & LSTM, GANs are the frameworks one needs to be aware of.

If you want to get understanding of algorithms and frameworks mentioned here, I have covered the *reference material in upcoming chapter, or you can refer following link for top 10 algorithms.*

U03.1: Top 10 ML Algorithms: *https://towardsdatascience.com/top-10-machine-learning-algorithms-for-data-science-cdb0400a25f9*

Domain Knowledge

This lack of domain knowledge, while perfectly understandable, can be a major barrier to data scientists. For one thing, it’s difficult to come up with project ideas in a domain that you don’t know much about. It can also be difficult to determine the type of data that may be helpful for a project — if you want to build a model to predict an outcome, you need to know what types of variables might be related to this outcome so you can make sure to gather the right data.

Knowing the domain is useful not only for figuring out projects and how to approach them but also for having rules of thumb for sanity checks on the data. Knowing how data is captured (is it hand-entered? Is it from machines that can give false readings for any number of reasons?) can help a data scientist with data cleaning and from going too far down the wrong path. It can also inform what true outliers are and which values might just be due to measurement error.

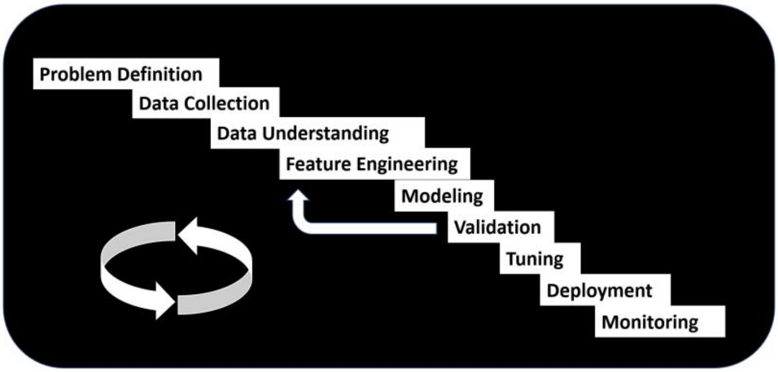
Often the most challenging part of building a machine learning model is feature engineering. Understanding variables and how they relate to an outcome is extremely important for this. Knowing the domain can help direct the data exploration and greatly speed (and enhance) the feature engineering process.

Once features are generated, knowing what relationships between variables are plausible help for basic sanity checks. Being able to glance at the outcome of a model and determine if they make sense goes a long way for quality assurance of any analytical work.

Finally, one of the biggest reasons a strong understanding of the data is important is because you have to interpret the results of analyses and modelling work.

Knowing what results are important and which are trivial is important for the presentation and communication of results. It's also important to know what results are actionable.

3.2 Process to Follow



Process to Follow
Working on Building Blocks

Problem Definition

The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

You'll often get ambiguous inputs from the people who have problems. You'll have to develop the intuition to turn scarce inputs into actionable outputs–and to ask the questions that nobody else is asking.

Data Collection

Once you've defined the problem, you'll need data to give you the insights needed to turn the problem around with a solution. This part of the process involves thinking through what data you'll need and finding ways to get that data, whether it's querying internal databases, or purchasing external data-sets.

Data Understanding

The difficulty here isn't coming up with ideas to test, it's coming up with ideas that are likely to turn into insights. You'll have a fixed deadline for your data science project, so you'll have to prioritize your questions.

You'll have to look at some of the most interesting patterns that can help explain why sales are reduced for this group. You might notice that they don't tend to be very active on social media, with few of them having Twitter or Facebook accounts. You might also notice that most of them are older than your general audience. From that you can begin to trace patterns you can analyze more deeply.

Feature Engineering

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process. Feature Engineering is in fact an art.

Modeling

Depending on the type of question that you're trying to answer, there are many modelling algorithms available. You run the selected algorithm/s on the training data to build the models.

Validation

Validation is a step used to evaluate the trained model on validation data. You use a series of steps competing for machine-learning algorithms along with the various associated tuning parameters that are geared toward answering the question of interest with the current data.

Tuning

Tuning an algorithm or machine learning technique can be simply thought of as a process which one goes through in which they optimize the parameters that impact the model in order to enable the algorithm to perform the best.

Deployment

After you have a set of models that perform well, you can operationalize them for other applications to consume. Depending on the business requirements, predictions are made either in real-time or on a batch basis. To deploy models, you expose them with an open API interface. The interface enables the model to be easily consumed from various applications.

U03.2: DS/AI Process: *https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview*

3.3 Tools to Master

SQL
python
R
TensorFlow
K Keras
PyTorch

Tools to Master
Working on Building Blocks

The list mentioned here is not exhaustive, it depends more on what kind of problem you are solving and in what tech stack you are working.

SQL

Structured Query Language (SQL) is a standard computer language for relational database management and data manipulation. SQL is used to query, insert, update and modify data. Most relational databases support SQL.

As data collection has increased exponentially, so has the need for people skilled at using and interacting with data; to be able to think critically, and provide insights to make better decisions and optimize their businesses. The skills necessary to be a good data scientist include being able to retrieve and work with data and to do that you need to be well versed in SQL, the standard language for communicating with database systems.

U03.3.1: SQL: *https://www.tutorialspoint.com/sql/*

R

R is a programming language and software environment for statistical analysis, graphics representation and reporting. In the world of data science, R is an increasingly popular language for a reason. It was built with statistical manipulation in mind, and there's an incredible ecosystem of packages for R that let you do amazing things — particularly in data visualization.

U03.3.2: R: *https://www.statmethods.net/r-tutorial/index.html*

Python

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is no-doubt the best-suited language for a Data Scientist. It is a free, flexible and powerful open-source language. Python cuts development time in half with its simple and easy to read syntax. With Python, you can perform data manipulation, analysis, and visualization. Python provides powerful libraries for Machine learning applications and other scientific computations.

U03.3.3: Python: *http://www.tutorialspoint.com/python/python_data_science.htm*

Tensorflow

Currently, the most famous deep learning library in the world is Google's TensorFlow. Google product uses machine learning in all of its products to improve the search engine, translation, image captioning or recommendations.

TensorFlow is the best library of all because it is built to be accessible to everyone. Tensorflow library incorporates different API to built at scale deep learning architecture like CNN or RNN. TensorFlow is based on graph computation; it allows the developer to visualize the construction of the neural network with Tensorboad. This tool is helpful to debug the program. Finally, Tensorflow is built to be deployed at scale. It runs on CPU and GPU.

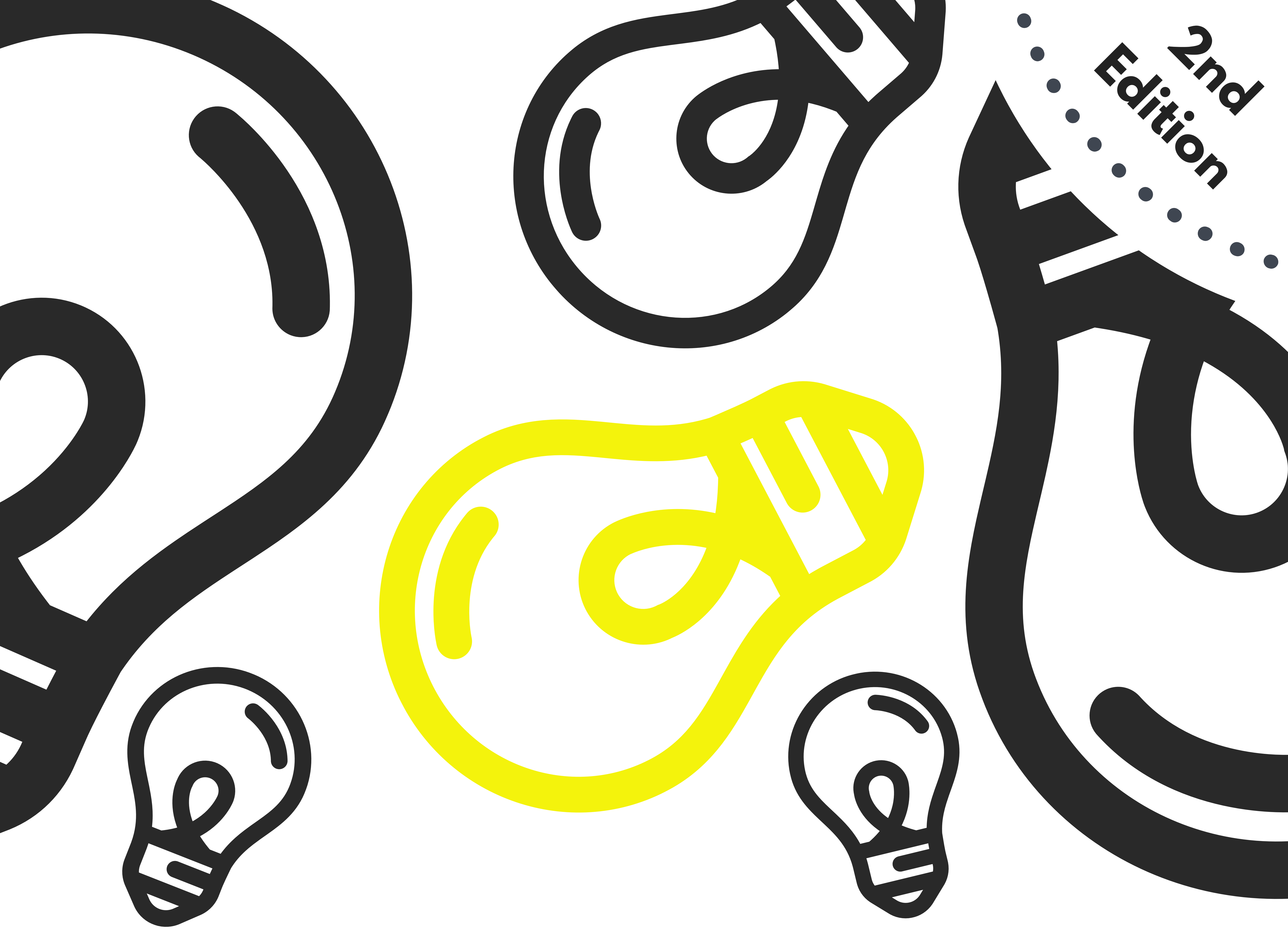
U03.3.4: Tensorflow: *https://www.guru99.com/tensorflow-tutorial.html*

Keras

Keras is a high-level neural networks API, capable of running on top of Tensorflow, Theano, and CNTK. It enables fast experimentation through a high level, user-friendly, modular and extensible API.

Keras allows for easy and fast prototyping (through user-friendliness, modularity, and extensibility). It supports both convolutional networks and recurrent networks, as well as combinations of the two. It runs seamlessly on CPU and GPU.

U03.3.5: Keras: *https://www.guru99.com/keras-tutorial.html*



2nd Edition

Artificial Intelligence

Self-Starter Handbook

BUILD YOUR OWN ROADMAP

Ankit Rathi

Coming Soon… 2nd Edition

with revised content & 3 more chapters…

ankitrathi.com

From a time around when AI field started picking up, every other day I get many questions from AI starters & enthusiasts on ‘How can I get into AI field?’. Over a while, I have improvised my response based on the follow-up questions they ask like:

- What is AI and why is it important?
- What is the difference between AI, ML, DL, DS, DM, BI?
- What an end-to-end AI project looks like?
- What are the roles in AI projects, who does what?
- What AI concepts & tools you need to learn?
- Which books, courses, channels etc you need to refer to?
- How to practice & build an AI portfolio?
- How to write a resume for an AI role?
- How to build a helpful network?
- How to search for the job?
- How to prepare for the interview?
- How to switch into an AI role (inside or outside)?
- How to lead an AI initiative in your organization?
- How to stay up-to-date in this ever-evolving field?

You can notice that these questions are not conceptual ones and there is no dedicated material to address these roadblocks. I thought why not to build a framework or a road-map for AI starters and enthusiasts so that I need not answer the same type of questions again and again. And that is when I started documenting what a starter or enthusiast need to do step by step in order to reach a level when he is ready to tackle any challenge thrown to him. My answer to the above questions in a structured way to help AI starters & enthusiasts is this book. This book covers the framework to launch your AI career in 11 chapters.

Ankit Rathi is a data & AI architect, published author & well-known speaker. His interest lies primarily in building end to end AI applications/products following best practices of Data Engineering and Architecture.

In his free time, he blogs about various topics on Data & AI field & tries to simplify it for starters & enthusiasts.

ankitrathi.com

₹ 499 | $ 9.99 | € 8.99
ISBN 978-1-64587-432-4