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

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

Chapter 4

Utilizing the Resources

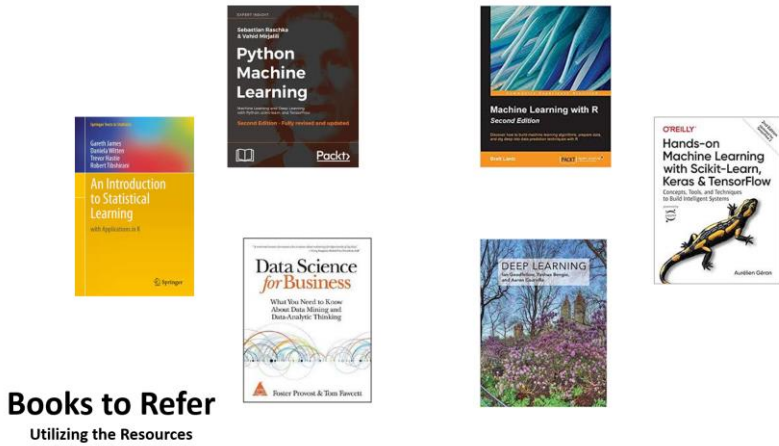
After getting to know DS/AI landscape and its building blocks in previous chapter, you may ask which are the resources to refer, how do I know if a book or course is worth to spend time and/or money?

This is not an exhaustive list by any means, but it is good enough to keep as your reference. You can build your own list of references once you get more awareness of the field.

This chapter is my attempt to make your task easier. I am listing down major quality resources (mostly free) here and also going to provide you with my view of these resources, which will help you to make an informed decision.

You need not go through each and every resource mentioned here, I would suggest you build the foundation first using a course or a book and keep other resources for your reference.

4.1 Books to refer



Machine Learning with R

This is an excellent book for the R starter who wants to apply ML to any kind of project. All the main ML models are presented, as well as different performance metrics, bagging, pruning, tuning, ensembling etc. Easy to scan through, many tips with fully-solved textbook problems. Certainly, a very good starting point if you plan to compete on Kaggle. If you already master both R and ML, this books is obviously not for you.

U04.1.1: Machine Learning with R: *https://www.goodreads.com/en/book/show/26033041-machine-learning-with-r---second-edition*

Python Machine Learning

This is a fantastic introductory book in machine learning with python. It provides enough background about the theory of each (covered)

technique followed by its python code. One nice thing about the book is that it starts implementing Neural Networks from scratch, providing the reader with the chance of truly understanding the key underlying techniques such as back-propagation. Even further, the book presents an efficient (and professional) way of coding in python, the key to data science.

U04.1.2: Python Machine Learning: *https://www.goodreads.com/book/show/25545994-python-machine-learning*

ISLR

The book explains the concepts of Statistical Learning from the very beginning. The core ideas such as bias-variance trade-off are deeply discussed and revisited in many problems. The included R examples are particularly helpful for beginners to learn R. The book also provides a brief, but concise description of functions' parameters for many related R packages. Compared to *The Elements of Statistical Learning*, it is easy for the reader to understand. It does a wonderful job of breaking things down complex concepts. If one wishes to learn more about a particular topic, I'd recommend The Element of Statistical Learning. These two pair nicely together.

U04.1.3: ISLR: *https://www.goodreads.com/book/show/17397466-an-introduction-to-statistical-learning*

Deep Learning

This is the book to read on deep learning. Written by luminaries in the field — if you've read any papers on deep learning, you must have heard about Goodfellow and Bengio before — and cutting through much of the BS surrounding the topic: like 'big data' before it, 'deep learning' is not something new and is not deserving of a special name. Networks with more hidden layers to detect higher-order features, networks of different

types chained together in order to play to their strengths, graphs of networks to represent a probabilistic model.

This is a theoretical book, but it can be read in tandem with Hands-On Machine Learning with Scikit-Learn and TensorFlow, almost chapter-for-chapter. The Scikit-Learn and Tensorflow example code, while only moderately interesting on its own, helps to clarify the purpose of many of the topics in the Goodfellow book.

U04.1.4: Deep Learning: *https://www.goodreads.com/en/book/show/24072897-deep-learning*

Hands-On Machine Learning with Scikit-Learn and TensorFlow

This book provides a great introduction to machine learning for both developer and non-developers. Authors suggest to just go through even if you don't understand math details. Highlights of this book are:

- Extraction of field expert knowledge is very important, you should know which model will serve better for the given solution. Luckily, a lot of models are available already from other scientists.
- Training data is the most important part, the more you have it the better. So if you can you should accumulate as much data as you can, preferably categorized, you may not still know how you will apply the accumulated data in the future but you will need it.
- Labelling training data is very important too, to train neural network you need to have at least thousands of labelled data samples, the more the better.
- Machine learning algorithms and neural networks are pretty common for years but the latest breakthrough is possible because of new optimization, new autoencoders (that may help to artificially

generate training data) allowing to do training faster and with fewer data.

- Machine learning is still pretty time and resources consuming process. To train a machine learning model you need to know how to tweak parameters and how to use different training approaches fitting the particular model.

The book demonstrates (including the code) different approaches using Scikit-Learn python package and also the TensorFlow.

U04.1.5: Hands-On Machine Learning with Scikit-Learn and TensorFlow: *https://www.goodreads.com/book/show/32899495-hands-on-machine-learning-with-scikit-learn-and-tensorflow*

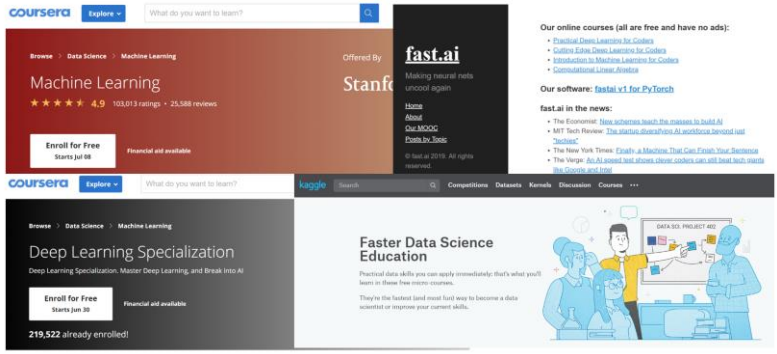
Data Science for Business

This is probably the most practical book to read if you are looking for an overview of data science. Either you know when terms like k-means and ROC curves are to be used or you have some context when you start digging deeper into how some of these algorithms are implemented. You will find it at the right level because there is just enough math to explain the fundamental concepts and make them stick in your head.

This isn't a book on implementing these concepts or a bunch of algorithms. This gives the book the advantage of being something you can refer to an intelligent manager or interested developer, and they can both get a lot out of it. And if they are interested in the next level of learning there are plenty of pointers. You will also find the chapter on presenting results through ROC curves, lift curves, etc. pretty interesting. It would be cool if this book had some more hands-on, but you can go to Kaggle and browse around the current and past competitions to apply what you learn here.

U04.1.6: Data Science for Business: *https://www.goodreads.com/book/show/17912916-data-science-for-business*

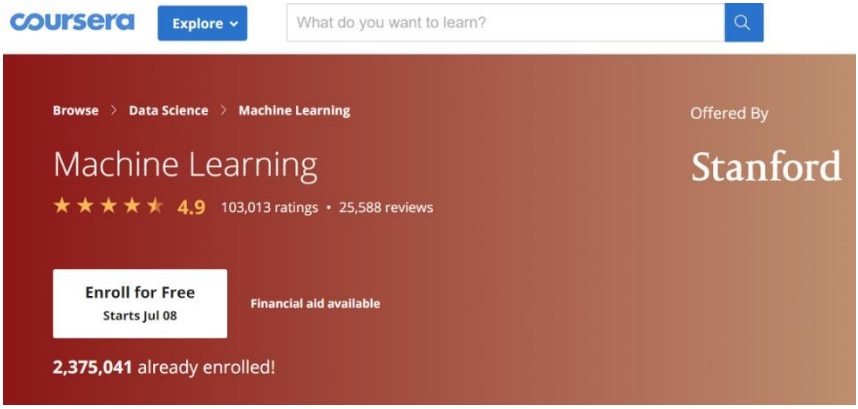
4.2 Courses to Attend



Courses to Attend

Utilizing the Resources

Machine Learning

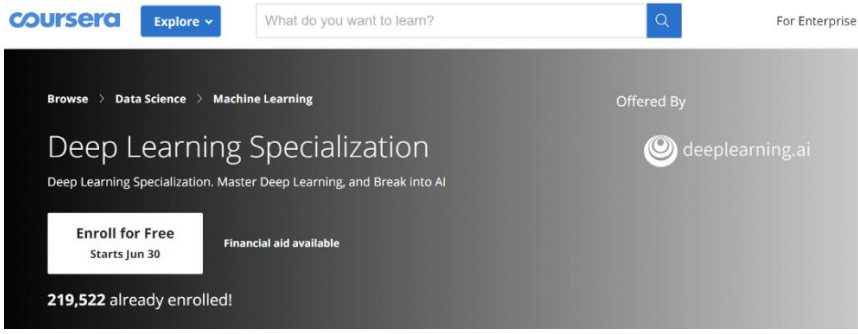


Machine Learning is one of the first programming MOOCs. Coursera put online by Coursera founder and Stanford Professor Andrew Ng. This course assumes that you have basic programming skills and you have some understanding of Linear Algebra. Knowledge of Statistics & Probability is not required though.

Andrew Ng does a good job explaining dense material and slides. The course gives you a lot of structure and direction for each homework, so it is generally pretty clear what you are supposed to do and how you are supposed to do it.

U04.2.1: Machine Learning: *https://www.coursera.org/learn/machine-learning*

Deep Learning

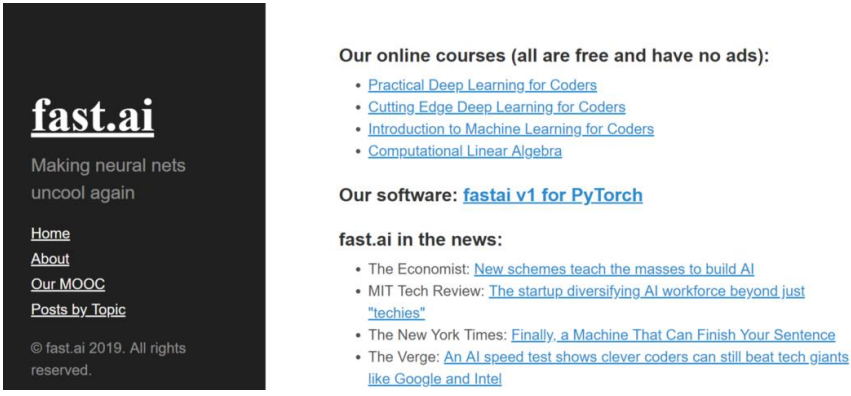


When you are rather new to the topic, you can learn a lot of doing the deeplearning.ai specialization. First and foremost, you learn the basic concepts of NN. How does a forward pass in simple sequential models look like, what's a backpropagation, and so on? I experienced this set of courses as a very time-effective way to learn the basics and worth more than all the tutorials, blog posts and talks, which I went through beforehand.

Doing this specialization is probably more than the first step into DL. I would say, each course is a single step in the right direction, so you end up with five steps in total. I think it builds a fundamental understanding of the field. But going further, you have to practice a lot and eventually it might be useful also to read more about the methodological background of DL variants. But doing the course work gets you started in a structured manner — which is worth a lot, especially in a field with so much buzz around it.

U04.2.2: Deep Learning: *https://www.deeplearning.ai/deep-learning-specialization/*

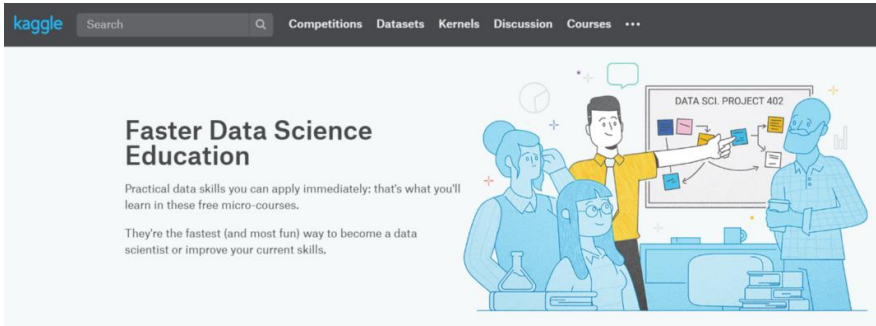
Fast AI



If your goal is to be able to learn about deep learning and apply what you've learned, the fast.ai course is a better bet. If you have the time, interleaving the deeplearning.ai and fast.ai courses is ideal — you get the practical experience, applicability, and audience interaction of fast.ai, along with the organised material and theoretical explanations of deeplearning.ai.

U04.2.3: Fast AI: *https://www.fast.ai/*

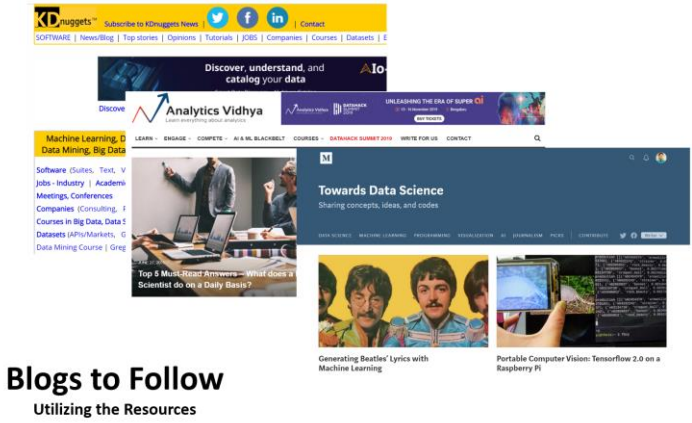
Kaggle Learn



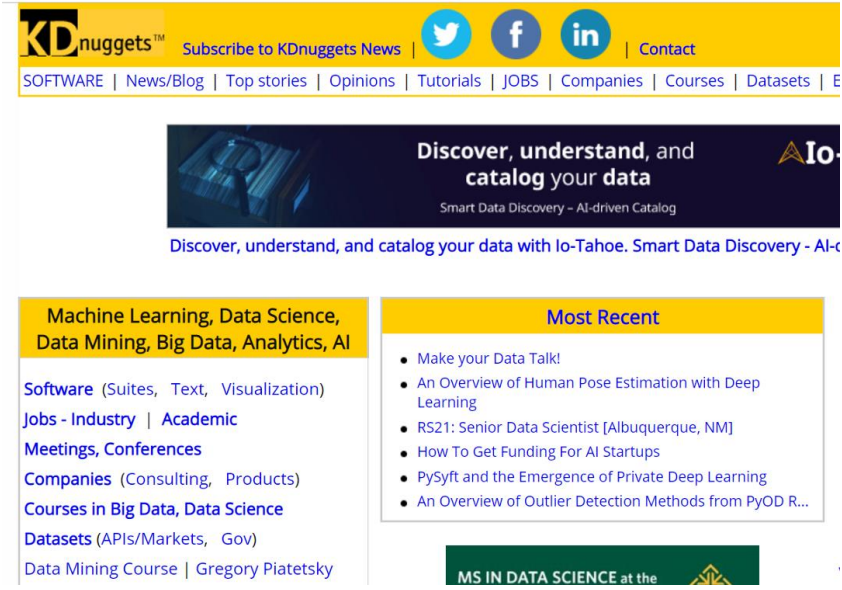
Practical data skills you can apply immediately: that's what you'll learn in these free micro-courses. They're the fastest (and most fun) way to become a data scientist or improve your current skills.

U04.2.4: Kaggle Learn: *https://www.kaggle.com/learn/overview*

4.3 Blogs to follow



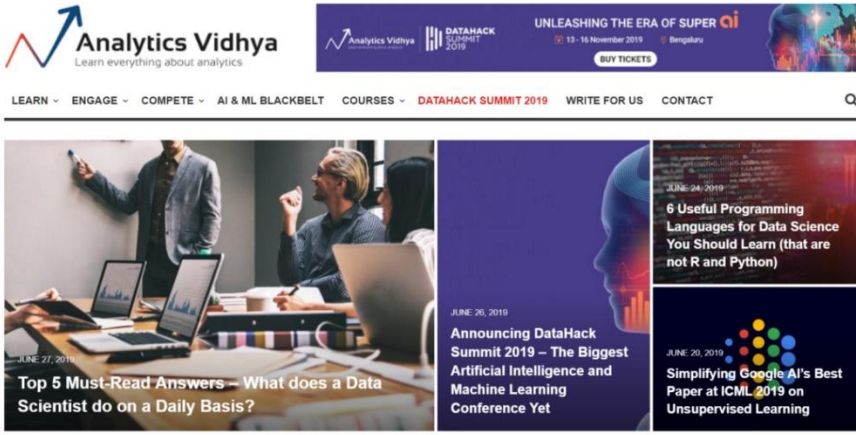
KD Nuggets



KDnuggets is a leading site on AI, Analytics, Big Data, Data Mining, Data Science, and Machine Learning and is edited by Gregory Piatetsky-Shapiro and Matthew Mayo. KDnuggets was founded in February of 1997. Before that, Gregory maintained an earlier version of this site, called Knowledge Discovery Mine, at GTE Labs (1994 to 1997).

U04.3.1: KD Nuggets: *https://www.kdnuggets.com/*

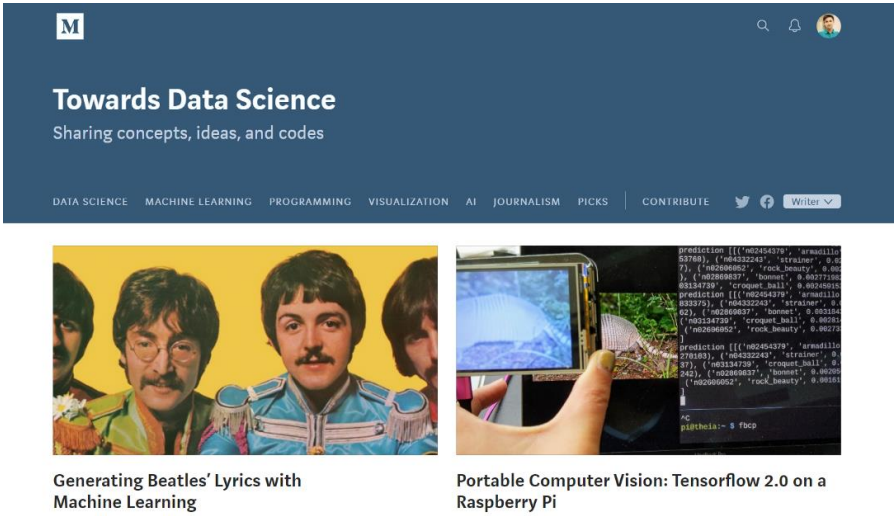
Analytics Vidhya



Analytics Vidhya provides a community-based knowledge portal for Analytics and Data Science professionals. The aim of the platform is to become a complete portal serving all knowledge and career needs of Data Science Professionals.

U04.3.2: Analytics Vidhya Blog: *https://www.analyticsvidhya.com/blog/*

Towards Data Science

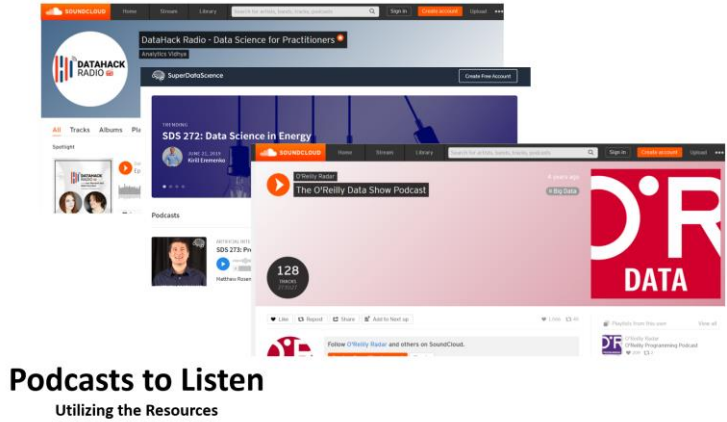


TDS joined Medium's vibrant community in October 2016. In the beginning, their goal was simply to gather good posts and distribute them to a broader audience. Just a few months later, they were pleased to see that they had a very fast-growing audience and many new contributors.

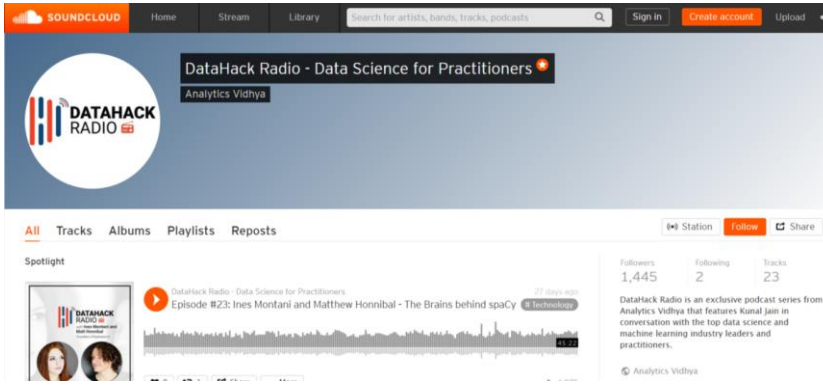
Today they are working with more than 10 Editorial Associates to prepare the most exciting content for our audience. They provide customized feedback to our contributors using Medium's private notes. This allows them to promote their latest articles across social media without the added complexity that they might encounter using another platform.

U04.3.3: Towards Data Science: *https://towardsdatascience.com/*

4.4 Podcasts to listen



Data Hack



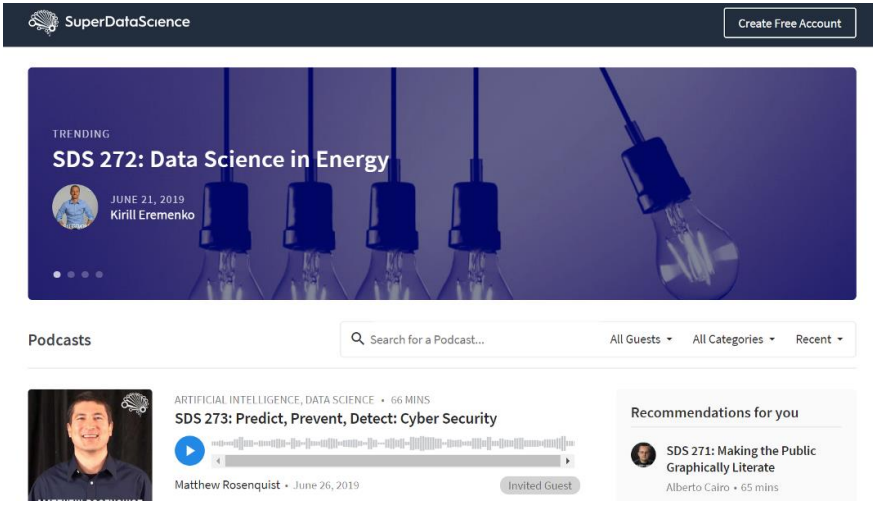
This is Analytics Vidhya's exclusive podcast series which will feature top leaders and practitioners in the data science and machine learning industry.

So in every episode of DataHack Radio, they bring you discussions with one such thought leader in the industry. They have discussions about

their journey, their learnings and plenty of other data science-related things.

U04.4.1: Data Hack Radio: *https://datahack.analyticsvidhya.com/*

Super Data Science

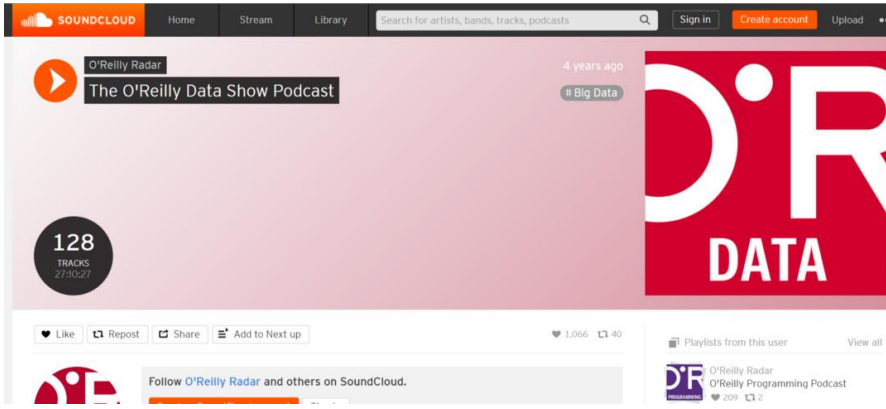


Kirill Eremenko is a Data Science coach and lifestyle entrepreneur. The goal of the *Super Data Science* podcast is to bring you the most inspiring Data Scientists and Analysts from around the World to help you build your successful career in Data Science.

Data is growing exponentially and so are salaries of those who work in analytics. This podcast can help you learn how to skyrocket your analytics career. Big Data, visualization, predictive modelling, forecasting, analysis, business processes, statistics, R, Python, SQL programming, tableau, machine learning, Hadoop, databases, data science MBAs, and all the analytics tools and skills that will help you better understand how to crush it in Data Science.

U04.4.2: Super Data Science: *https://www.superdatascience.com/podcast*

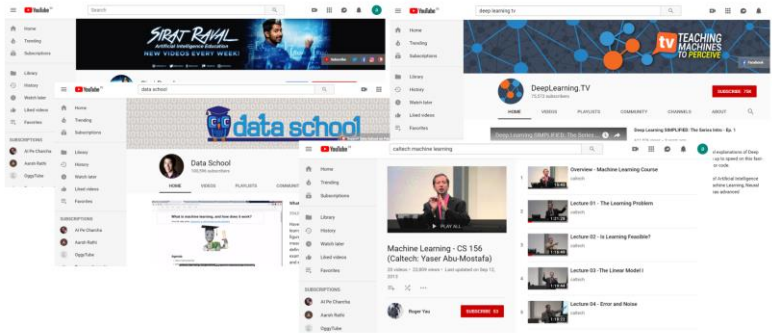
The O'Reilly Data Show Podcast



Known as the father of all other data shows, "the O'Reilly Data Show" features Ben Lorica, O'Reilly Media's chief data scientist. Lorica conducts interviews with other experts about big data and data science current affairs. While it does get technical and may not be the best place for a beginner to start, it provides interesting insights into the future of the data science industry.

U04.4.3: The O'Reilly Data Show Podcast: *https://www.oreilly.com/topics/oreilly-data-show-podcast*

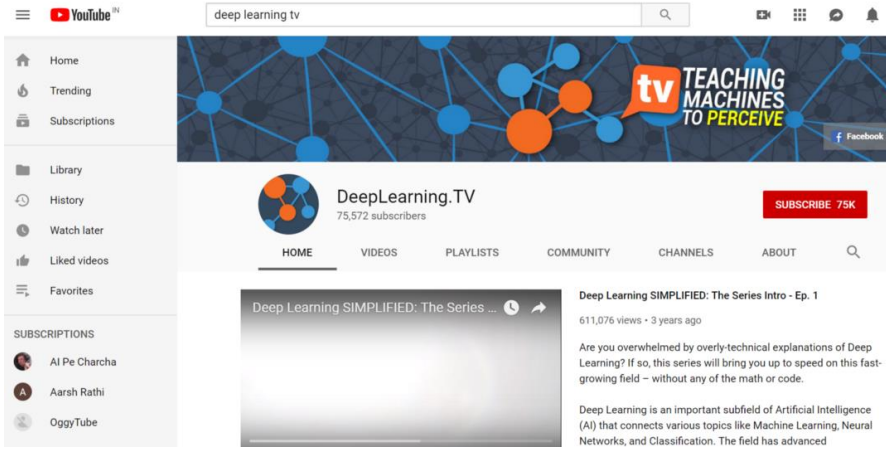
4.5 YouTube Channels



YouTube Channels

Utilizing the Resources

DeepLearning.TV

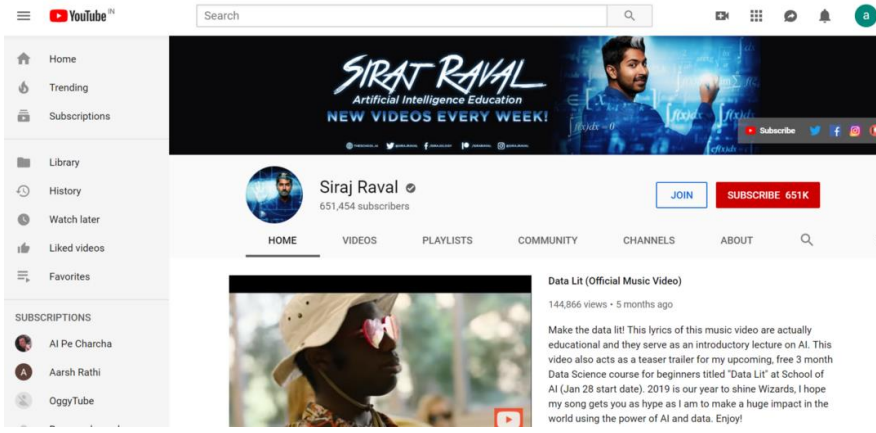


DeepLearning.TV is all about Deep Learning, the field of study that teaches machines to perceive the world. Starting with a series that simplifies Deep Learning, the channel features topics such as How To's, reviews of software libraries and applications, and interviews with key

individuals in the field. Through a series of concept videos showcasing the intuition behind every Deep Learning method, they show you that Deep Learning is actually simpler than you think. Their goal is to improve your understanding of the topic so that you can better utilize Deep Learning in your own projects. They provide a window into the cutting edge of Deep Learning and bring you up to speed on what's currently happening in the field.

U04.5.1: DeepLearning.TV:
https://www.youtube.com/playlist?list=PLjJh1vlSEYgvGod9wWiydumYl8hOXixNu

Sirajology



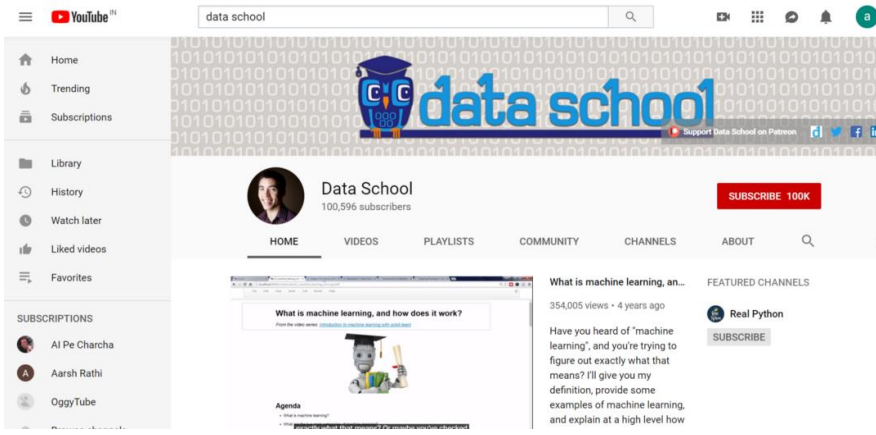
Your host here is Siraj. He is on a warpath to inspire and educate developers to build Artificial Intelligence. Games, music, chatbots, art, he teaches you how to make it all yourself. This is the fastest-growing AI community in the world. Their mission: Solve AI. Use it to benefit humanity.

He is an AI Researcher, his latest paper is here — *https://drive.google.com/file/d/0BwUv84lNDk72Q1qzaXqwR2U3U2NWVlZSOFk4amZIRmV1QXI0/view*

He is also a Data Scientist, AI Educator, Rapper, Author, and Director of the School of AI (www.theschool.ai)

U04.5.2: Sirajology: *https://www.youtube.com/channel/UCWN3xxRkmTPmbKwht9FuE5A*

Data School

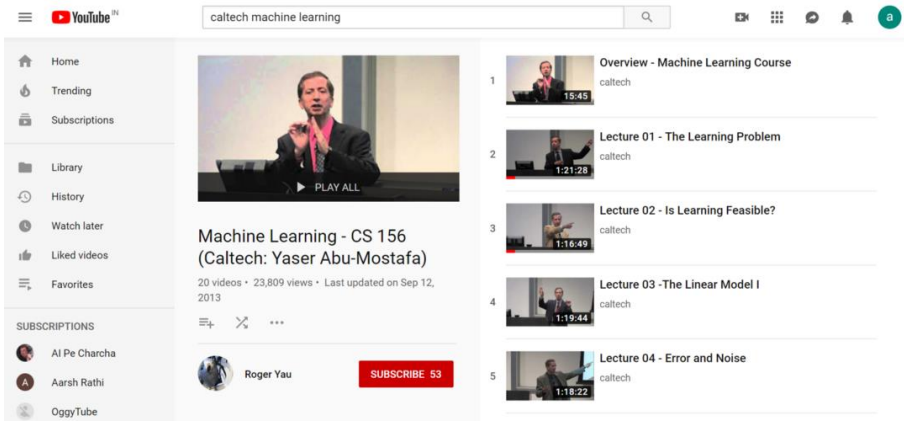


Are you trying to learn data science so that you can get your first data science job? You're probably confused about what you're "supposed" to learn, and then you have the hardest time actually finding lessons you can understand! Data School focuses you on the topics you need to master first, and offers in-depth tutorials that you can understand regardless of your educational background.

Your host here is Kevin Markham, and he is the founder of Data School. He has taught data science using the Python programming language to hundreds of students in the classroom, and hundreds of thousands of students (like you) online. Finding the right teacher was so important to his data science education, and so he sincerely hopes that he can be the right data science teacher for you.

U04.5.3: Data School: *https://www.youtube.com/dataschool*

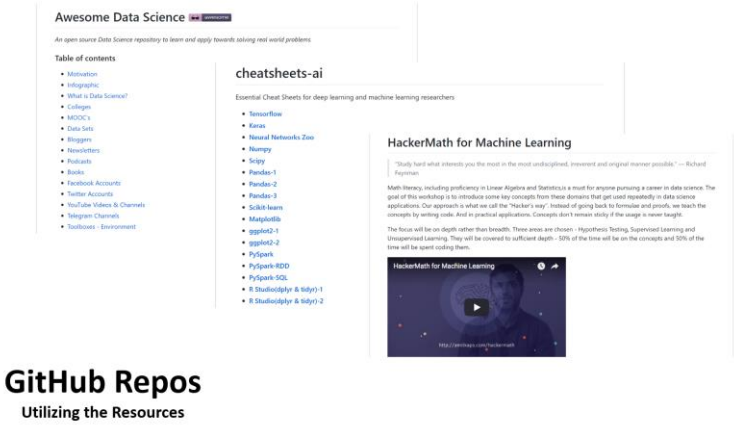
Caltech Machine Learning



This is an introductory course by Caltech Professor Yaser Abu-Mostafa on machine learning that covers the basic theory, algorithms, and applications. Machine learning (ML) enables computational systems to adaptively improve their performance with experience accumulated from the observed data. ML techniques are widely applied in engineering, science, finance, and commerce to build systems for which we do not have a full mathematical specification (and that covers a lot of systems). The course balances theory and practice and covers the mathematical as well as the heuristic aspects.

U04.5.4: CalTech Machine Learning: *https://www.youtube.com/playlist?list=PLD63A284B7615313A*

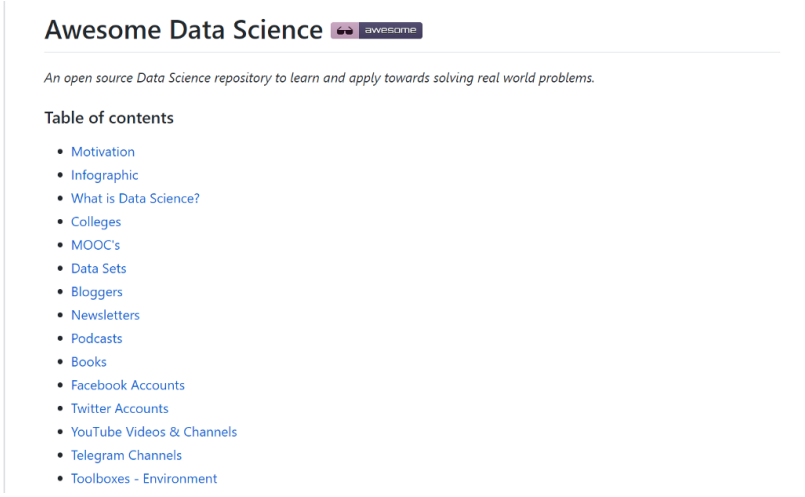
4.6 GitHub Repos



GitHub Repos

Utilizing the Resources

Awesome Data Science



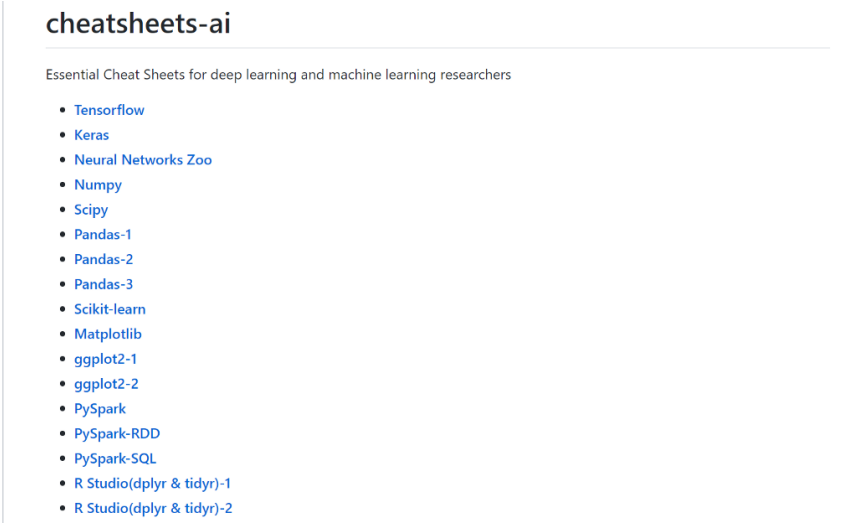
This Repo answer the questions, “What is Data Science and what should you study to learn Data Science?” An awesome Data Science repository to learn and apply for real-world problems.

As the aggregator says, “Our favourite data scientist is Clare Corthell. She is an expert in data-related systems and a hacker and has been working on a company as a data scientist. Clare’s blog. This website helps you to understand the exact way to study as a professional data scientist.”

“Secondly, Our favourite programming language is Python nowadays for Data Science. Python’s — Pandas library has full functionality for collecting and analyzing data. We use Anaconda to play with data and to create applications.”

U04.6.1: Awesome Data Science: *https://github.com/bulutyazilim/awesome-datascience*

Essential Cheat Sheets for Machine Learning and Deep Learning Engineers

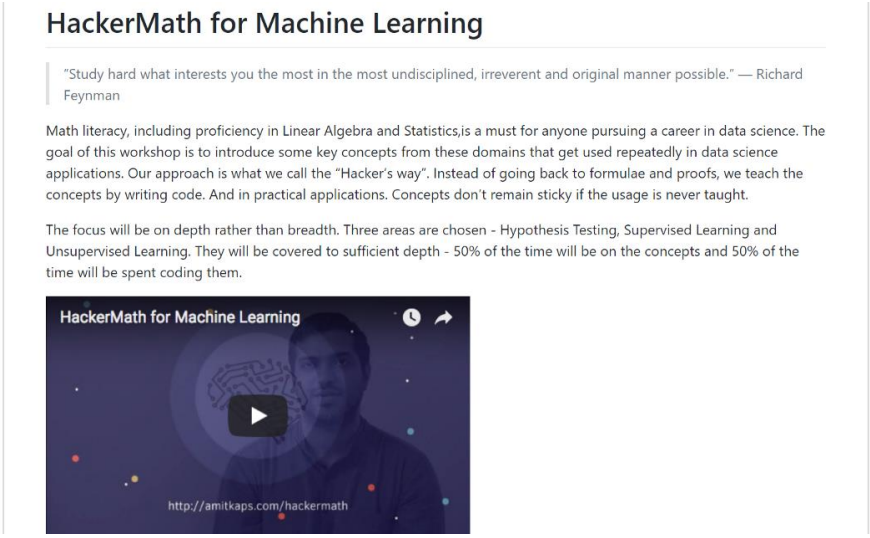


Machine learning is complex. For newbies, starting to learn machine learning can be painful if they don’t have the right resources to learn from. Most of the machine learning libraries are difficult to understand

and the learning curve can be a bit frustrating. Kailash Ahirwar has created a repository on Github (cheatsheets-ai) containing cheatsheets for different machine learning frameworks, gathered from different sources. Have a look at the Github repository, also, contribute cheat sheets if you have any.

U04.6.2: Cheatsheets-AI: *https://github.com/kailashahirwar/cheatsheets-ai*

HackerMath for Machine Learning

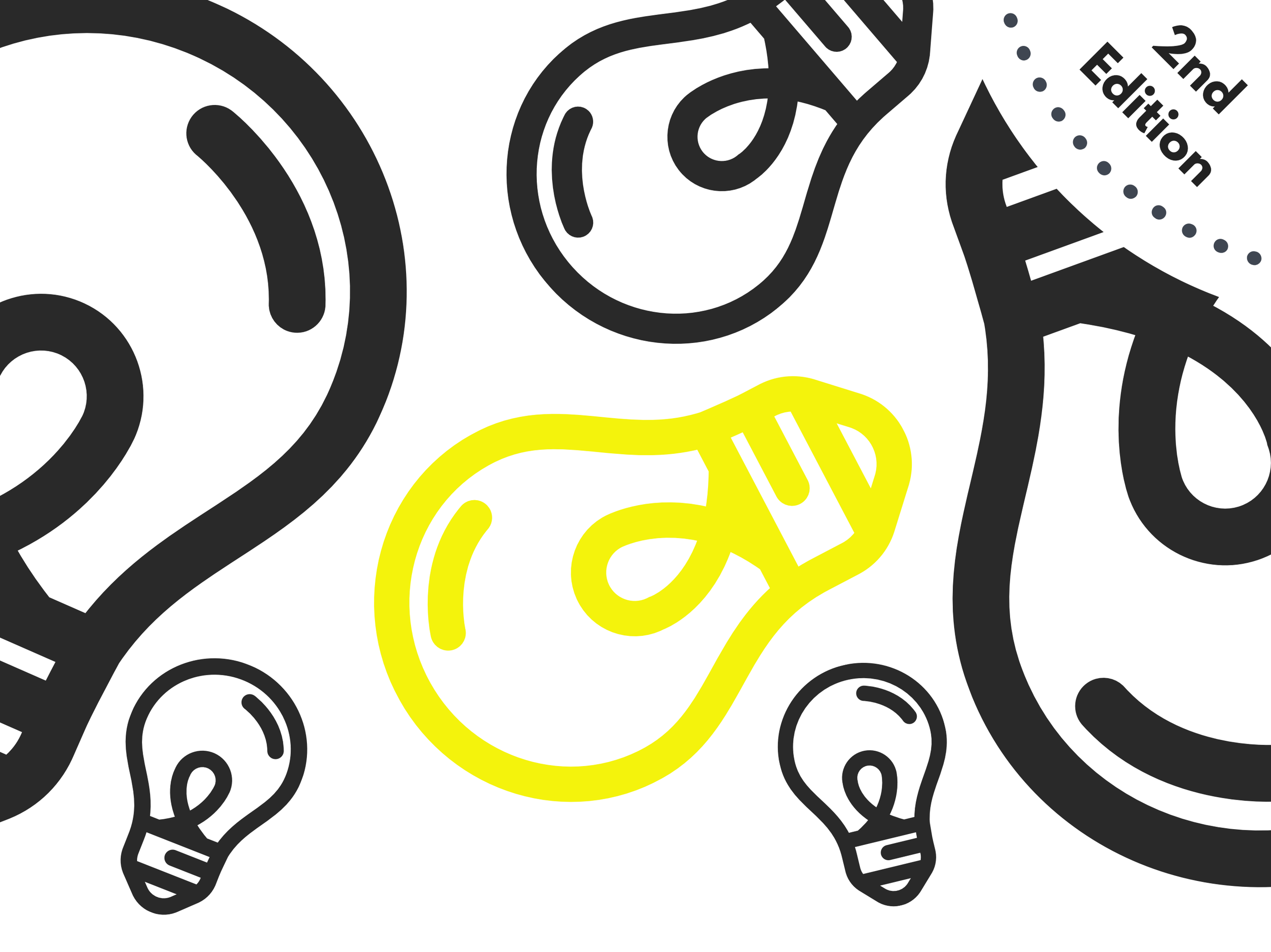


Math literacy, including proficiency in Linear Algebra and Statistics, is a must for anyone pursuing a career in data science. The goal of this workshop is to introduce some key concepts from these domains that get used repeatedly in data science applications.

As outlined by Amit Kapoor, “Our approach is what we call the ‘Hacker’s way’. Instead of going back to formulae and proofs, we teach the concepts by writing code. And in practical applications. Concepts don’t remain sticky if the usage is never taught.”

The focus here is on depth rather than breadth. Three areas are chosen — Hypothesis Testing, Supervised Learning and Unsupervised Learning. They are covered to sufficient depth — 50% of the time on the concepts and 50% of the time spent coding them.

U04.6.3: HackerMath for Machine Learning: *https://github.com/amitkaps/hackermath*



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