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

ar 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

Launch

DS/AI: Self-Starter Kit

Build Your Own Roadmap

Chapter 5

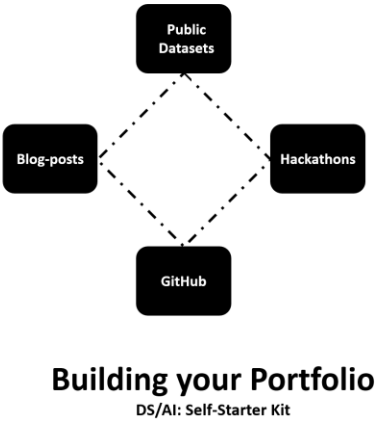
Building your Portfolio

This chapter talks about how you can build your DS/AI portfolio. Lets first understand, why a portfolio is important in DS/AI field?

Besides the benefit of learning by making a portfolio, a portfolio is important as it can help get you employment.

For the purpose of this article, let’s define a portfolio as public evidence of your DS/AI skills.

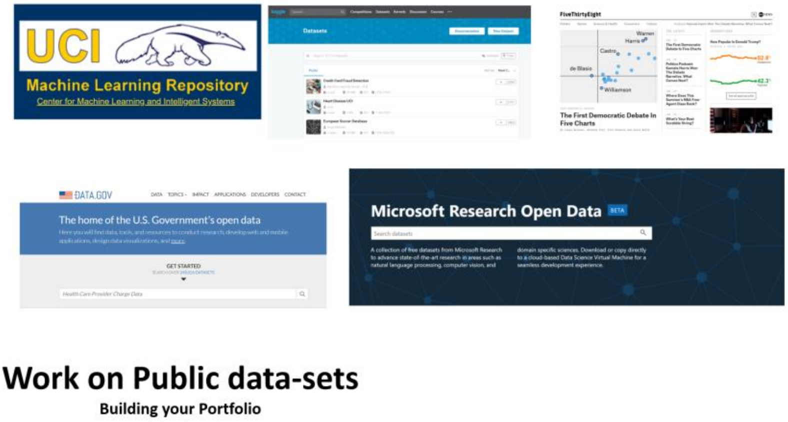
People often forget that software engineers and data scientists also google their issues. If these same people have their problems solved by reading your public work, they might think better of you and reach out to you.



Building your Portfolio

DS/AI: Self-Starter Kit

5.1 Work on Public data-sets


UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
Datasets
Credit Card Fraud Detection
Heart Disease UCI
European Soccer Database
FiveThirtyEight
Warren
Harris
Castro
de Blasio
Williamson
The First Democratic Debate In Five Charts
DATA.GOV
DATA
TOPICS
IMPACT
APPLICATIONS
DEVELOPERS
CONTACT
GET STARTED
Health Care Provider Charge Data
Microsoft Research Open Data
BETA
Search datasets

The home of the U.S. Government’s open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and domain specific sciences. Download or copy directly to a cloud-based Data Science Virtual Machine for a seamless development experience.

Work on Public data-sets

Building your Portfolio

You can gain more DS/AI skills by working on prediction problems rather than getting stuck in endless learning loop.

But you will not get a project to work on from day one of your learning. Still, there are platforms where you can apply and learn DS/AI.

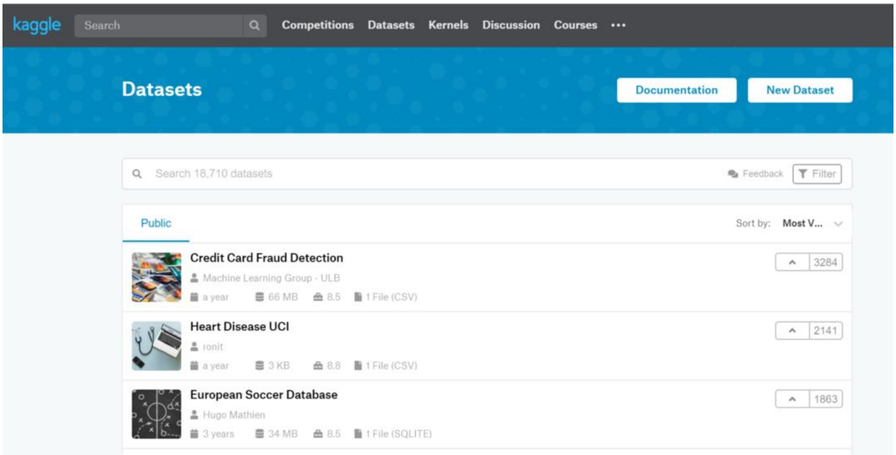
UCI ML

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

The UCI Machine Learning Repository is a collection of data-sets that are used by the machine learning community for the analysis of machine learning algorithms. The archive was created as an FTP archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data-sets. As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science. The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman, and this project is in collaboration with Rexa.info at the University of Massachusetts Amherst. Funding support from the National Science Foundation is gratefully acknowledged.

U05.1.1: UCI ML: *https://archive.ics.uci.edu/ml/index.php*

Kaggle Datasets



Credit Card Fraud Detection
Machine Learning Group - ULB
a year 66 MB 8.5 1 File (CSV) 3284

Heart Disease UCI
ronit
a year 3 KB 8.8 1 File (CSV) 2141

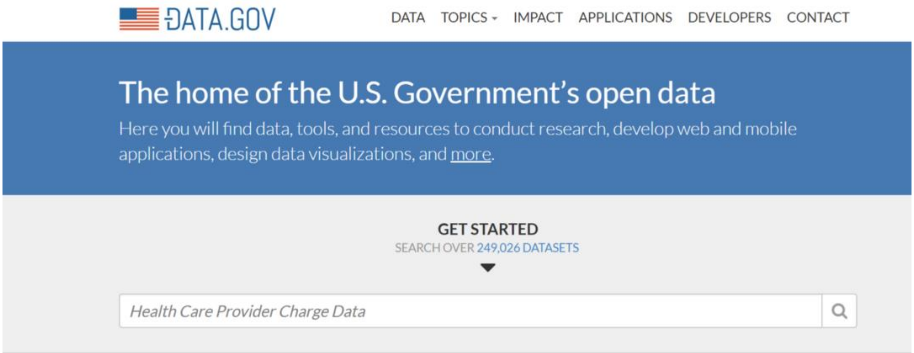
European Soccer Database
Hugo Mathien
3 years 34 MB 8.5 1 File (SQLITE) 1863

Kaggle is where many data scientists spend their nights and weekends. It's a crowd-sourced platform to attract, nurture, train and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over half a million active members from 190+ countries and it receives close to 150K submissions per month. Started from Melbourne, Australia Kaggle moved to Silicon Valley in 2011, ultimately been acquired by the Google in March of 2017. Kaggle is the number one stop for data science enthusiasts all around the world who compete for prizes and boost their Kaggle rankings. There are only a handful of Kaggle Grandmasters in the world to this date.

Do you know that most data scientists are only theorists and rarely get a chance to practice before being employed in the real-world? Kaggle solves this problem by giving data science enthusiasts a platform to interact and compete in solving real-life problems. The experience you get on Kaggle is invaluable in preparing you to understand what goes into finding feasible solutions for big data.

U05.1.2: Kaggle Datasets: *https://www.kaggle.com/datasets*

Data.Gov



The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

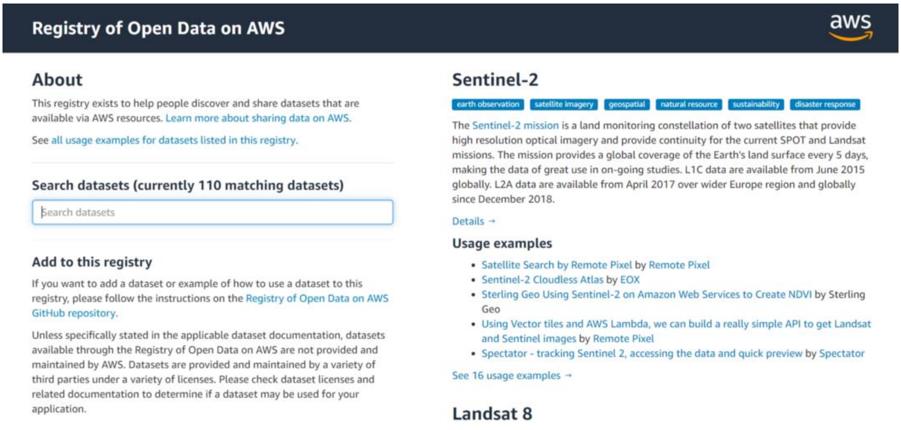
SEARCH OVER 249,026 DATASETS

Health Care Provider Charge Data

This is the home of the U.S. Government’s open data. Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more. Data.gov is managed and hosted by the U.S. General Services Administration, Technology Transformation Service. Data.gov is powered by two open source applications: CKAN and WordPress, and it is developed publicly on GitHub. Learn how you can contribute to Data.gov and these larger open source projects here.

U05.1.3: Data.Gov: *https://www.data.gov/*

Amazon Data-sets



Registry of Open Data on AWS

About

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS.

See all usage examples for datasets listed in this registry.

Search datasets (currently 110 matching datasets)

Search datasets

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

Sentinel-2

earth observation satellite imagery geospatial natural resource sustainability disaster response

The Sentinel-2 mission is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region and globally since December 2018.

Details →

Usage examples

- Satellite Search by Remote Pixel by Remote Pixel
- Sentinel-2 Cloudless Atlas by EOX
- Sterling Geo Using Sentinel-2 on Amazon Web Services to Create NDVI by Sterling Geo
- Using Vector tiles and AWS Lambda, we can build a really simple API to get Landsat and Sentinel images by Remote Pixel
- Spectator - tracking Sentinel 2, accessing the data and quick preview by Spectator

See 16 usage examples →

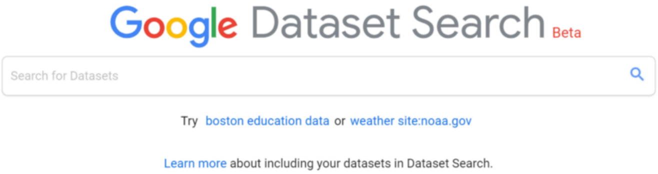
Landsat 8

This source contains many datasets in different fields such as Public Transport, Ecological Resources, Satellite Images, etc. It also has a search box to help you find the dataset you are looking for and it also has dataset description and Usage examples for all datasets which are very informative and easy to use!

The datasets are stored in Amazon Web Services (AWS) resources such as Amazon S3 — A highly scalable object storage service in the Cloud. If you are using AWS for machine learning experimentation and development, that will be handy as the transfer of the datasets will be very quick because it is local to the AWS network.

U05.1.4: Amazon Datasets: *https://registry.opendata.aws/*

Google’s Datasets Search Engine



Google Dataset Search Beta

Search for Datasets

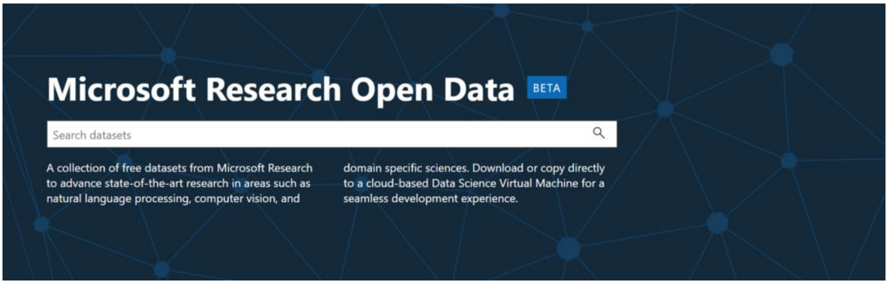
Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.

In late 2018, Google did what they do best and launched another great service. It is a toolbox that can search for datasets by name. Their aim is to unify tens of thousands of different repositories for datasets and make that data discoverable.

U05.1.5: Google Dataset Search: *https://toolbox.google.com/datasetsearch*

Microsoft Data-sets



Microsoft Research Open Data BETA

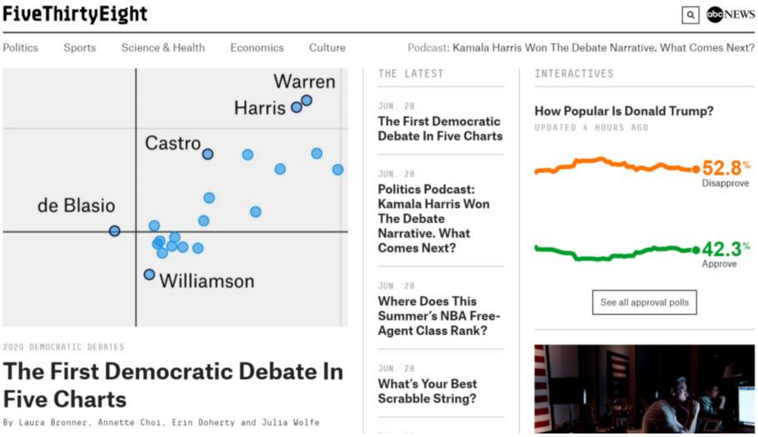
Search datasets

A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and domain specific sciences. Download or copy directly to a cloud-based Data Science Virtual Machine for a seamless development experience.

In July 2018, Microsoft along with the external research community announced the launch of “Microsoft Research Open Data”. It contains a data repository in the cloud dedicated to facilitating collaboration across the global research community. It offers a bunch of curated datasets that were used in published research studies.

U05.1.6: Microsoft Datasets: *https://msropendata.com/*

FiveThirtyEight



Podcast: Kamala Harris Won The Debate Narrative. What Comes Next?

2020 DEMOCRATIC DEBATES

The First Democratic Debate In Five Charts

By Laura Bronner, Annette Choi, Erin Doherty and Julia Wolfe

THE LATEST

JUN. 28
The First Democratic Debate In Five Charts

JUN. 28
Politics Podcast: Kamala Harris Won The Debate Narrative. What Comes Next?

JUN. 28
Where Does This Summer's NBA Free-Agent Class Rank?

JUN. 28
What's Your Best Scrabble String?

INTERACTIVES

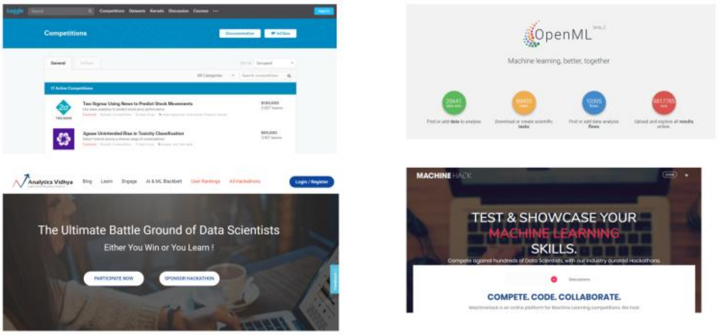
How Popular Is Donald Trump?
UPDATED 4 HOURS AGO

FiveThirtyEight, sometimes rendered as 538, is a website that focuses on opinion poll analysis, politics, economics and sports blogging. The website, which takes its name from the number of electors in the United States electoral college, was founded on March 7, 2008, as a polling aggregation website with a blog created by analyst Nate Silver.

You can find the data and code behind some of the popular articles and graphics here. You can use it to check others' work and to create stories and visualizations of your own.

U05.1.7: FiveThirtyEight: *https://fivethirtyeight.com/*

5.2 Participate in Hackathons

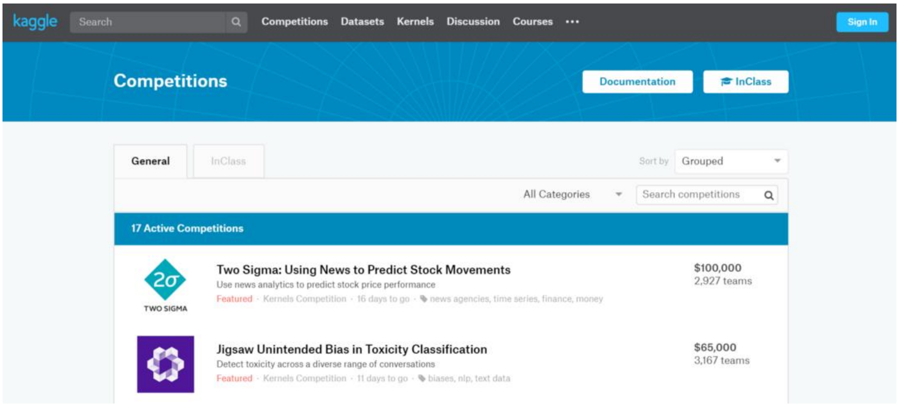

Competitions
Documentation
InClass
General
Two Sigma: Using News to Predict Stock Movements
Jigsaw Unintended Bias in Toxicity Classification
OpenML
Machine learning, better, together
20641
Find or add data to analyse
Download or create scientific tasks
10395
Find or add data analysis flows
Upload and explore all results online.
Analytics Vidhya
Blog
Learn
Engage
AI & ML Blackbelt
User Rankings
All Hackathons
Login / Register
The Ultimate Battle Ground of Data Scientists
Either You Win or You Learn !
PARTICIPATE NOW
SPONSOR HACKATHON
MACHINE HACK
TEST & SHOWCASE YOUR
MACHINE LEARNING
SKILLS.
Compete against hundreds of Data Scientists, with our industry curated Hackathons.
COMPETE. CODE. COLLABORATE.

Participate in Hackathons

Building your Portfolio

Participating in DS/AI competitions is one of the most frequent paths taken by data scientists, while it doesn't dish you all the challenges, it can help you to build your exploratory, modelling & cross-validation skills. You can also learn from fellow competitors about their approaches once the competition is over.

Kaggle Competitions



17 Active Competitions

Two Sigma: Using News to Predict Stock Movements
Use news analytics to predict stock price performance
Featured · Kernels Competition · 16 days to go · news agencies, time series, finance, money
$100,000
2,927 teams

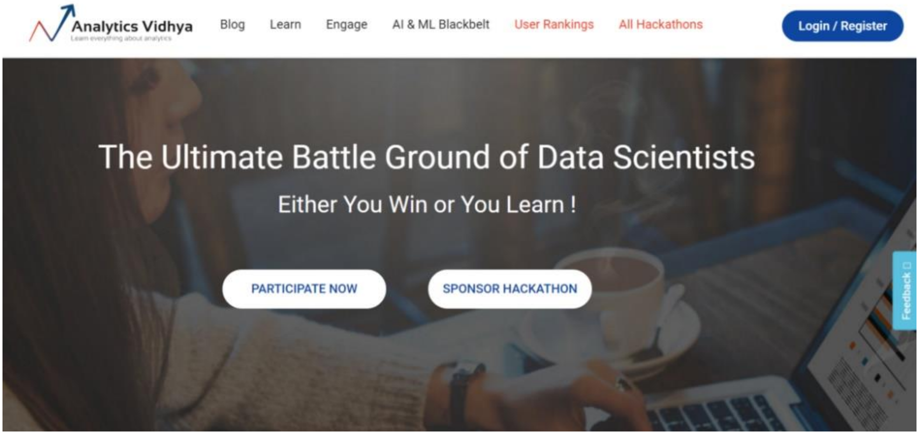
Jigsaw Unintended Bias in Toxicity Classification
Detect toxicity across a diverse range of conversations
Featured · Kernels Competition · 11 days to go · biases, nlp, text data
$65,000
3,167 teams

Kaggle runs a variety of different kinds of competitions, each featuring problems from different domains and have different difficulties. Before you start, navigate to the Competitions listing. It lists all of the currently active competitions.

If you click on a specific Competition in the listing, you will go to the Competition’s homepage.

U05.2.1: Kaggle Competitions: *https://www.kaggle.com/competitions*

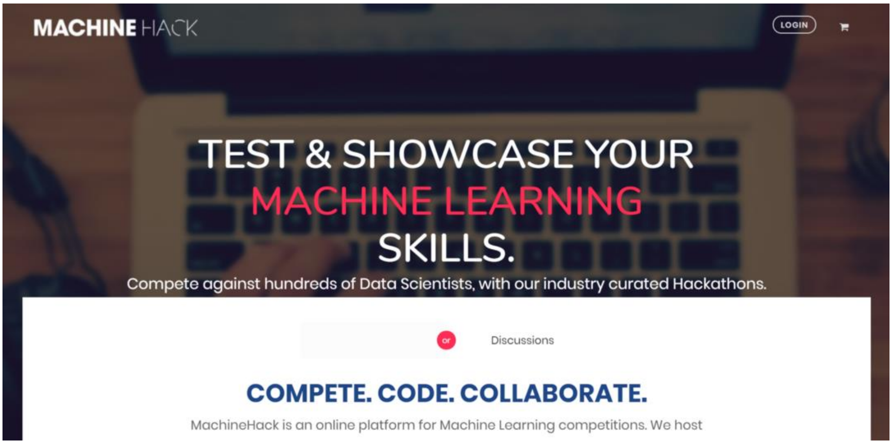
DataHack by AnalyticsVidhya



AnalyticsVidhya Data Hack is also a platform where you can compete with the best in the world on real-life data science problems. You can learn by working on real-world problems. You can also upskill yourself and get hired in the listed companies. You can showcase your expertise and get hired in top firms. If you happen to be at the top of competitions, you can also win lucrative prizes.

U05.2.2: Data Hack: *https://datahack.analyticsvidhya.com/*

Machine Hack



TEST & SHOWCASE YOUR MACHINE LEARNING SKILLS.

Compete against hundreds of Data Scientists, with our industry curated Hackathons.

COMPETE. CODE. COLLABORATE.

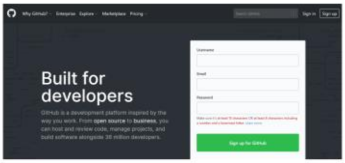
MachineHack is an online platform for Machine Learning competitions. We host

MachineHack is an online platform for Machine Learning competitions. They host the toughest business problems that can now find solutions using Machine Learning & Data Science techniques. Companies can hire better data scientists, the can discover & evaluate talented data scientists.

Just like Kaggle & DataHack, you can enrol in competitions here and help host solve their business problem. In return, you get near real-world project experience, you can learn from fellow competitors once the competition is over.

U05.2.3: Machine Hack: *https://www.machinehack.com/*

5.3 Publish on Git-hub



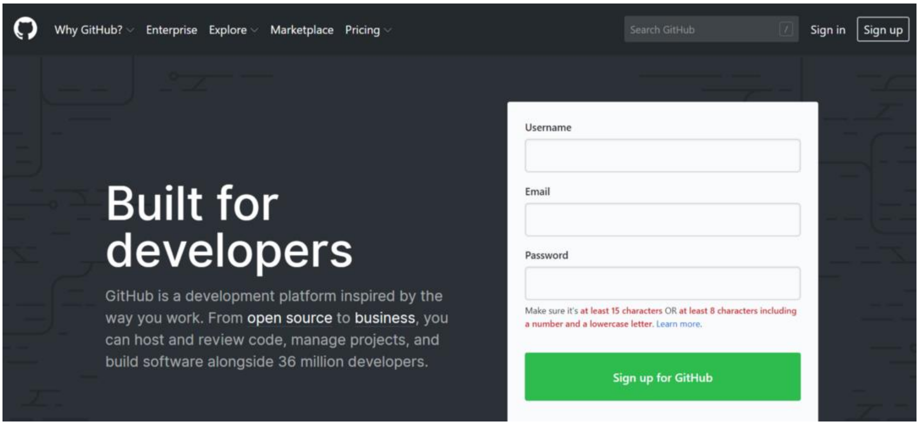
GitHub Pages
Websites for you and your projects.
Education from the experts

Publish on Github

Building your Portfolio

GitHub is a powerful platform for software development, but at its heart, it's about empowering people like you by helping you learn from other developers, build the software that matters to you, and propel yourself to the next stage of your life as a software developer.

Understand GitHub



Built for developers

GitHub is a development platform inspired by the way you work. From open source to business, you can host and review code, manage projects, and build software alongside 36 million developers.

Username

Email

Password

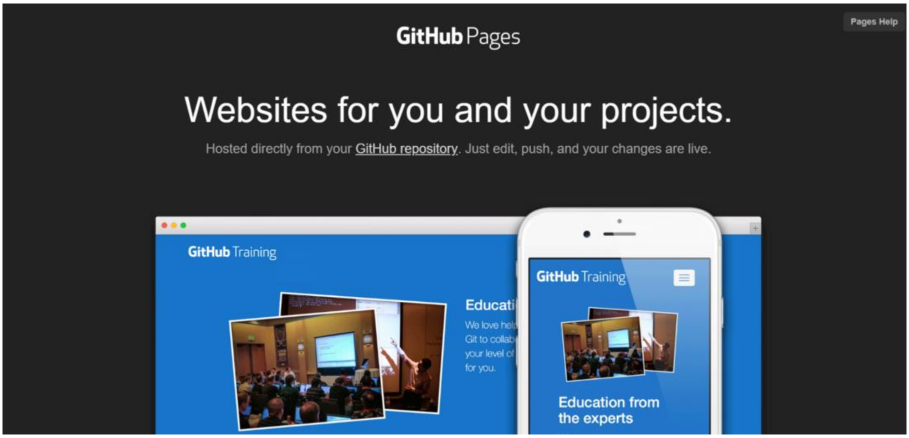
Make sure it's at least 15 characters OR at least 8 characters including a number and a lowercase letter. Learn more.

GitHub is a code hosting platform for version control and collaboration. It lets you and others work together on projects from anywhere.

In order to work on GitHub, you need to learn essentials like repositories, branches, commits, and Pull Requests. You'll create your own Hello World repository and learn GitHub's Pull Request workflow, a popular way to create and review code.

U05.3.1: GitHub: *https://github.com/*

GitHub Pages



Websites for you and your projects.

Hosted directly from your GitHub repository. Just edit, push, and your changes are live.

GitHub Pages are public webpages hosted and easily published through GitHub. The quickest way to get up and running is by using the Jekyll Theme Chooser to load a pre-made theme. You can then modify your GitHub Pages' content and style remotely via the web or locally on your computer.

U05.3.2: GitHub Pages: *https://pages.github.com/*

5.4 Write a Blog

Writing blogs is an effective way to showcase your expertise and skills. You can write what you have learnt recently, any interesting problem you have solved or worked on any project.

Writing an engaging blog-post is an art in itself, here are few tips to write and promote your blog-posts.

Take notes for ideas

Start by writing down ideas as they occur to you. Make it a habit and keep doing it consistently by installing a note-taking app (like Keep, EverNote etc) on your mobile device.

Ideas occur to us all the time. You need a way to capture them when they do so that you can turn them into a great blog-post in the future.

Build a simple outline

It is an essential step to develop an easy-to-follow outline before you sit down to write a blog-post.

Once you've picked a topic to write about, from the list of ideas that you've written down, create an outline. The outline contains a heading, introduction, major points you want to write about and conclusion.

To get the juices flowing, you should actually write the introduction and the conclusion first, then add a list of things that you'll cover in the body.

Start with a story

Entertainment is the biggest factor in engaging your audience. If you're just about to start a blog, keep this at in your mind.

Stories engage people in and help clear the doubts. You are able to develop a scene which people can relate to.

Become a memorable writer by integrating stories into your blog-posts. It doesn't have to be your own story, you can tell interesting stories about others.

Solve common problems

Consistent writing is one of the easiest ways to become a better writer. The question is, what should you write about? As a beginner, write blog-posts that answer questions.

Look for the problems that are common in your field, what most of the people are struggling with. Research about that topic, try to explore the problem and its possible solution.

Learn & Share

When I write a blog-post, I read a lot about the subject. On the web and in real life, there are too many questions with too few answers.

Many a time, you will end up learning yourself in an attempt to write the post on a certain topic.

Read other great writers

The truth is that if you don't read great writers, you don't really know how to do it and that successful blog that you dream of will evade you.

I've learned that I get a better education from studying authors' best work than I do from waiting for a piece of advice from them.

Mentor Others

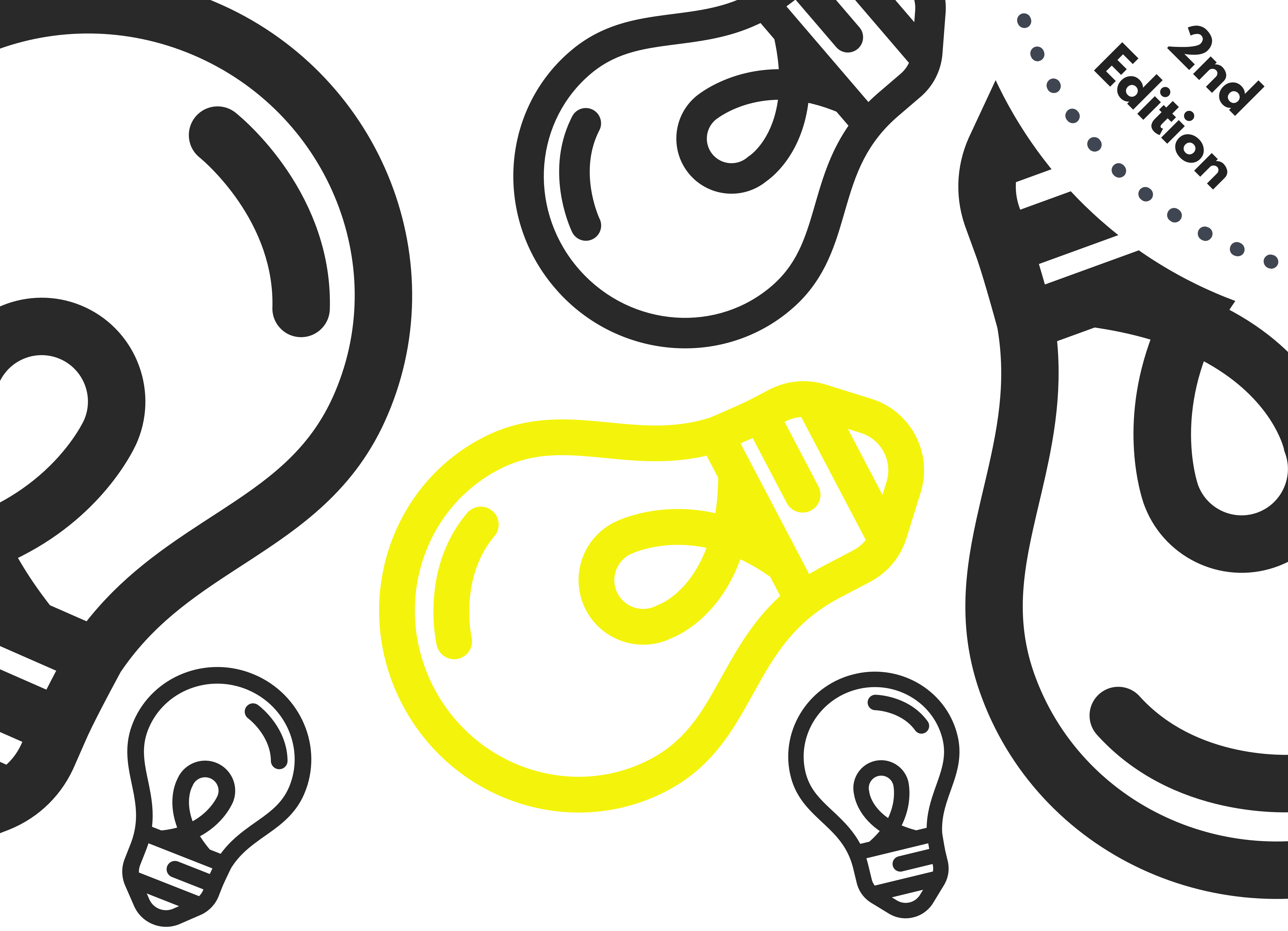
As part of being a successful and well-rounded data scientist, giving back can be a rewarding and beneficial aspect. Becoming a mentor, or mentoring those who want to follow in your steps of being a data scientist can sharpen your expertise and credentials.

Build a Personal Brand

Building a brand is about giving yourself more opportunities to help and connect with people in your industry. And one of the best ways to build a brand is through blogging.

A blog is a hub for your advice. It also has the added benefit of helping you rank on search engines.

I hope that reading this inspires at least a few of you who want to become a data scientist and want to get better day by day by following the above-mentioned approach.



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