

# Career in Data science

Dr. Sishir Kalita  
Data Scientist  
Armsoftech.air  
Chennai

*sishirkalita@outlook.com*

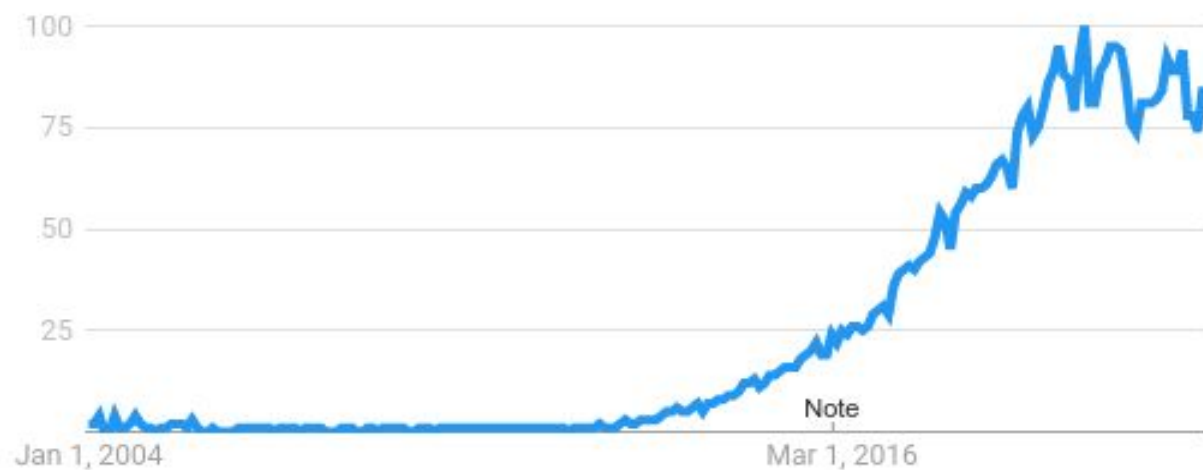
# Outline

1. What is data science?
2. Why data science is so popular today?
3. Components in data science
4. CRISP-DM Methodology
5. Data science job roles
6. Model-driven vs Data-driven approach
7. Current status - Kaggle report 2021
8. How to start?

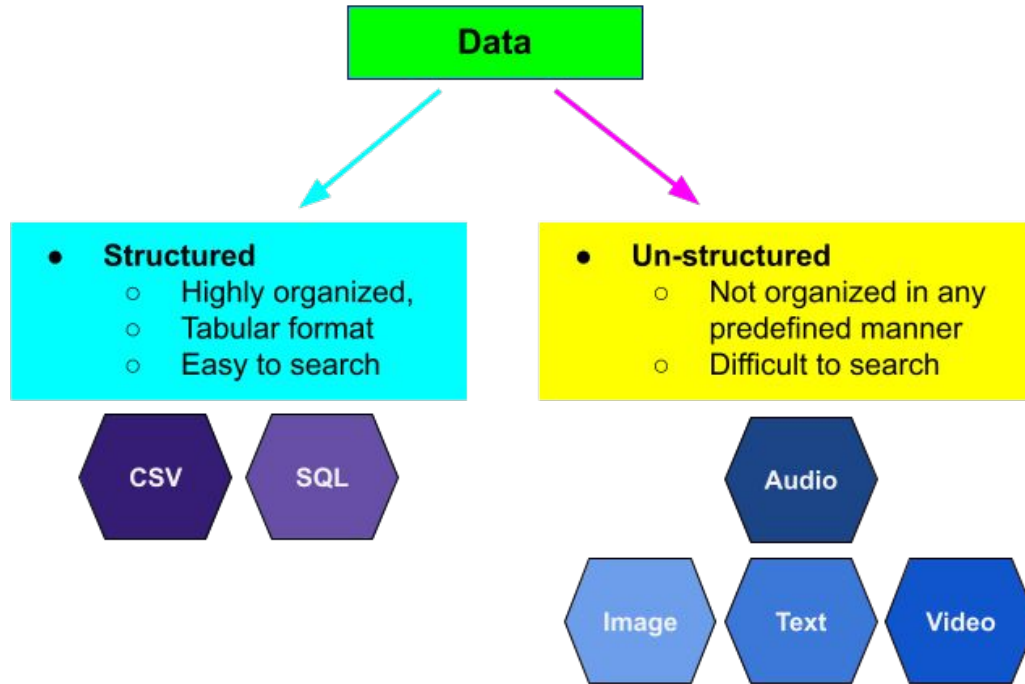
## Interest over time

Google Trends

● Data science



Worldwide. 1/1/04 - 2/2/22. Web Search.



*80% of enterprise data is unstructured data [1]*

# What is data science?

- Assortment of different tasks to uncover useful intelligence from data for an application



- Roughly speaking, data science is therefore simply the act of approaching the information extraction, processing and communication task in a scientific way

“The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”

– Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

**The goal is to turn data  
into information, and information  
into insight.**

**- Carly Fiorina  
ex CEO of Hewlett-Packard**



# Why is data science so popular?

- Digital revolution
- Internet of Things
- Social media
- E-commerce
- Big data revolution

*“Idea is to get insights from the data”*



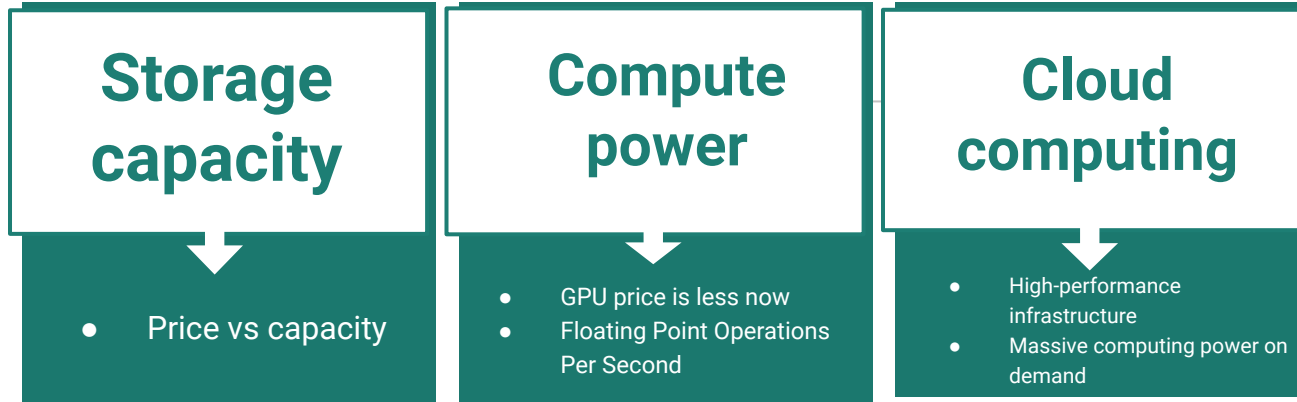




**There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.**

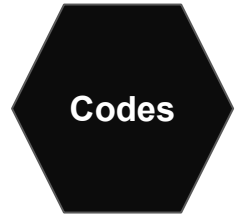
**- Eric Schmidt  
Executive Chairman of Google**

# Why is data science so popular?

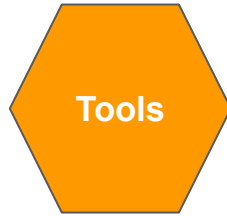


*Devices have become very powerful and cheaper*

# Why is data science so popular?

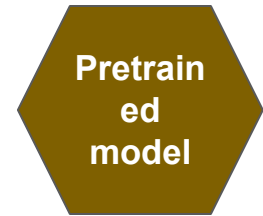


GitHub



data.world

AI4Bharat



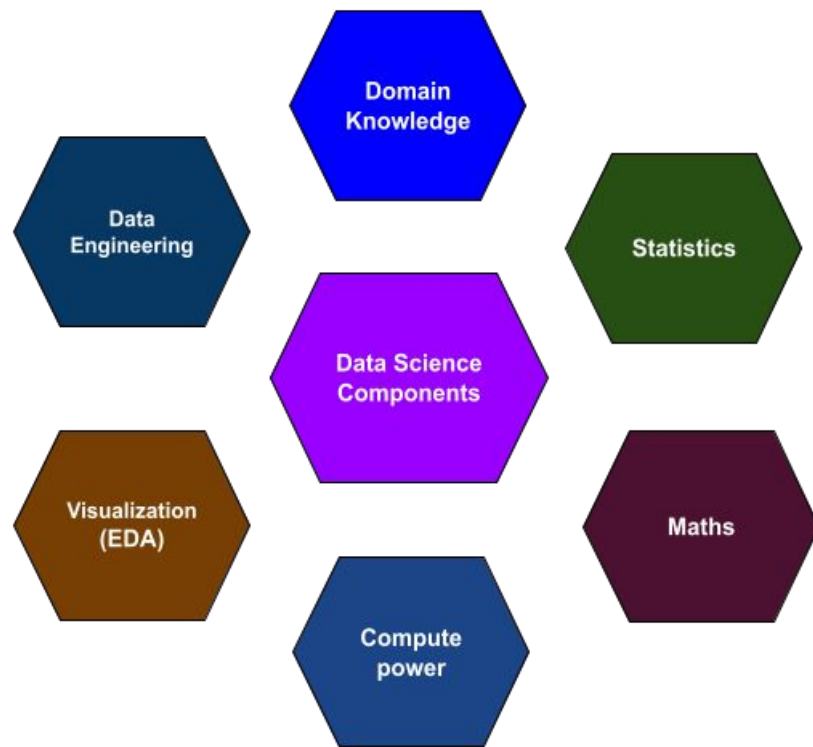
AI4Bharat



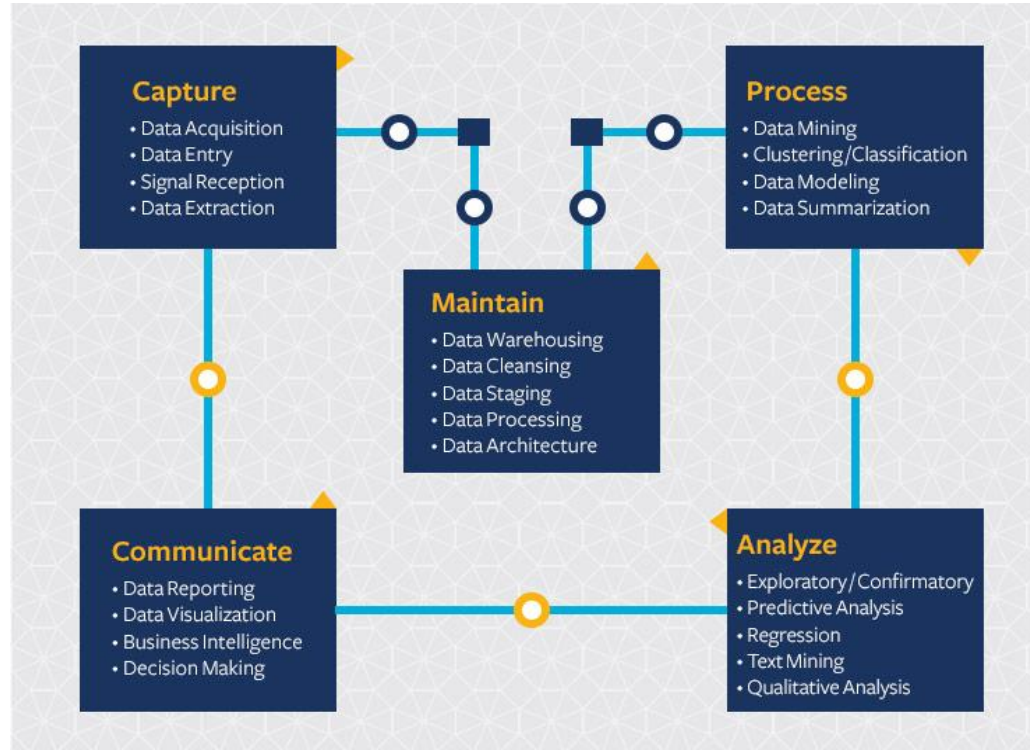
HUGGING FACE

*Open source movement is democratizing data science*

## Data Science Components






# The Data Science Life Cycle



[Source: ischoolonline.berkeley.edu](https://ischoolonline.berkeley.edu)

# Where will I get the data?

- Your own customized data
-  dataset search ([datasetsearch.research.google.com/](https://datasetsearch.research.google.com/)) - general
-  Datasets ([www.kaggle.com/datasets](https://www.kaggle.com/datasets)) - general
- Commonvoice Mozilla ([commonvoice.mozilla.org/en/datasets](https://commonvoice.mozilla.org/en/datasets)) - speech
- OpenSLR ([openslr.org/resources.php](https://openslr.org/resources.php)) - speech
- [National Platform for Language Technology](https://nlp.cba.hawaii.edu/NationalPlatformforLanguageTechnology/) - speech + NLP
- Linguistic Data Consortium ([catalog.ldc.upenn.edu](https://catalog.ldc.upenn.edu)) - speech
-  **Datasets** ([huggingface.co/datasets](https://huggingface.co/datasets)) - speech + NLP
- [EkStep-ULCA-Speech-Dataset](https://ekstep-ulca-speech-dataset.github.io/) - speech
- [UC Irvine Machine Learning Repository](https://nlp.cs.ucirvine.edu/)
- [Voice datasets](https://voice-datasets.github.io/) - speech

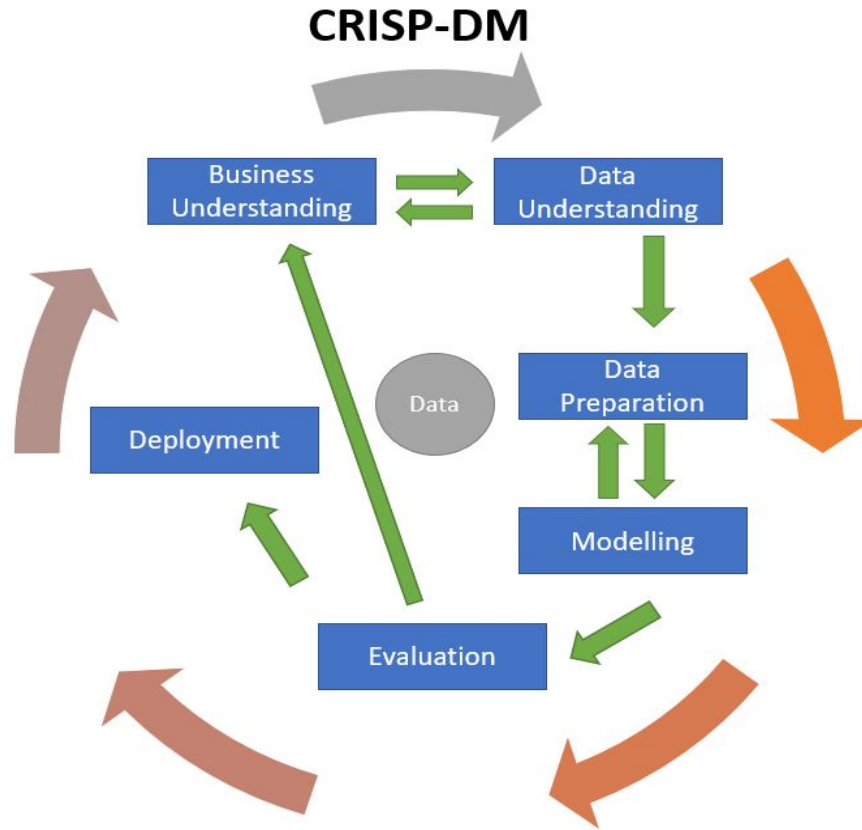
# CRISP-DM - Methodology For the Data Science Project

- **CRoss Industry Standard Process for Data Mining (CRISP-DM)**
- Six Cycles:
  - Business understanding – What does the business need?
  - Data understanding – What data do we have/need? Is it clean?
  - Data preparation – How do we organize the data for modeling?
  - Modeling – What modeling techniques should we apply?
  - Evaluation – Which model best meets the business objectives?
  - Deployment – How do stakeholders access the results?

# Iterative design and deployment approach

- First, build a minimal viable product (MVP) and establish the benchmark. Here, we will integrate all the basic components of the product and deploy that for live testing.
- Revise the first version of the system based on the feedback from the business team and improve.





[Image Caption: towardsdatascience.com](https://towardsdatascience.com)

## Business understanding:

- What are the business objectives and requirements? With priority?
- Why can data science achieve those objectives?
- What is the story to tell?
- How do we define the success metric?
- Any ethical issues in using the data?
- What have other industries achieved?

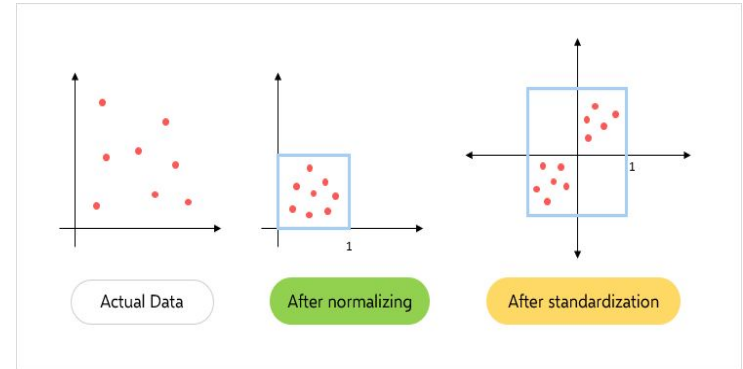


## Data understanding:

- What is the source of data?
- Does new data to be collected?
- Which data is relevant for the project?
- If we have the data, what is the quality OR quantity?

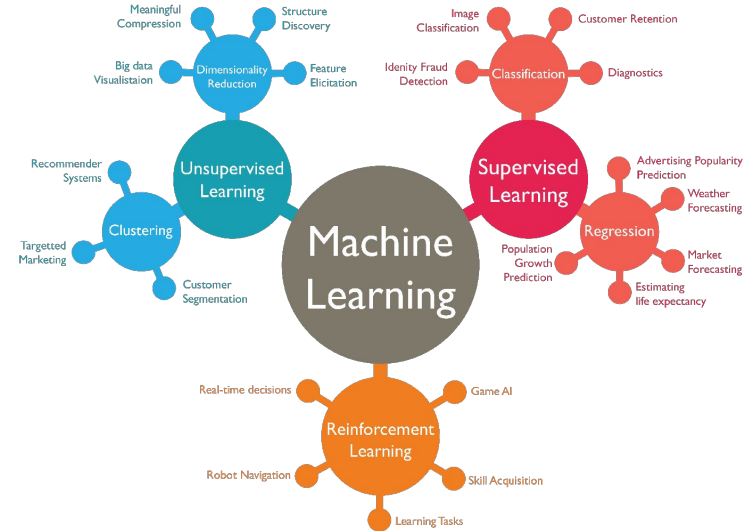
## Data preparation:

- Different data formats - JSON, CSV, WAV
- If we have the data:
  - Is it annotated? What is the ground truth (label)
- How can data be extracted, transformed, loaded?
- Formatting the data as per our requirement.
- Prepare train set and a clean dev set
- Store the data for analysis



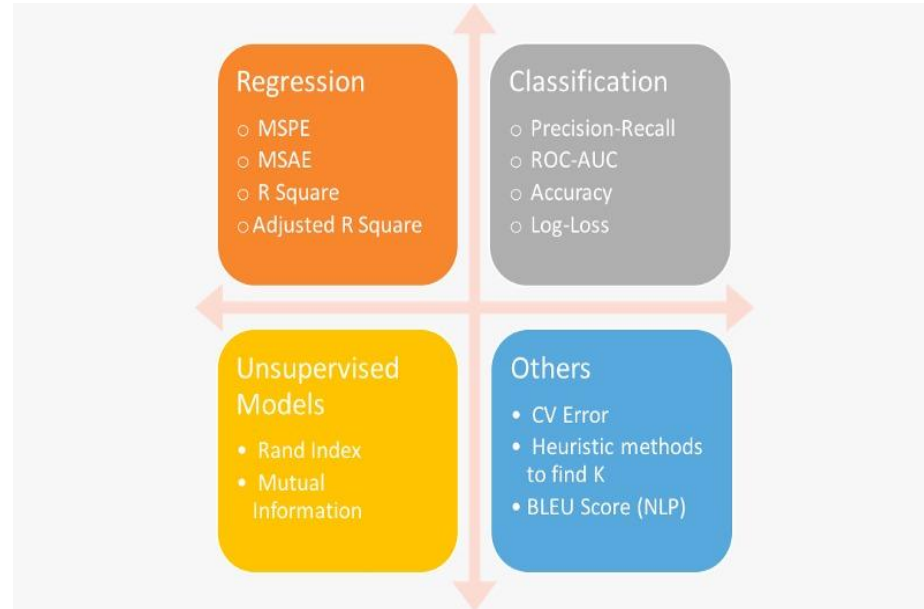
# Modeling:

- What assumption to make for the models?
- ML or DL?
- Check if the data is clean/sufficient for the modeling
  - Needs to do some statistical analysis to validate
- Statistical significant features?
- Which framework to use for building the models?
- Compute resources?



# Evaluation:

- What will be the performance metric to evaluate?
- Define the testing. Does the model work correctly on the test data?
- Does the model achieve the business objectives?
- Meet performance requirements?
- Unbiased or robustness
- Ways to improve the model

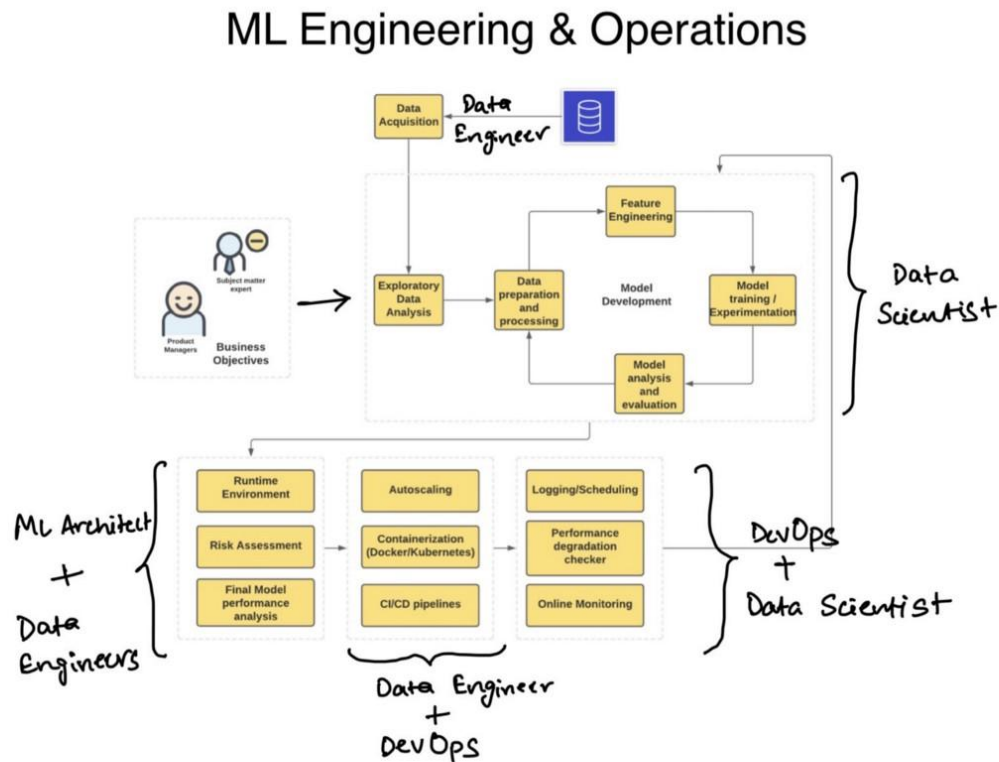


# Deployment:

- Where to be deployed?
- HW/SW stack to be used for deployment
- Inspection of latency/size of the model. Is it fine?
- Privacy requirement?
- Meet user expectations?

# Job Roles in Data Science Industry

- Data scientist
- Data analyst
- Machine learning engineer
- Data engineer
- Research scientist
- Speech scientist
- AI Scientist
- NLP Engineer
- Business intelligence





# Model centric to data centric approach

- Data Science System = Code (model/algorithm) + Data

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

**Improving the code vs. the data**

Source: [MLOps: From Model-centric to Data-centric AI](#)

# Rise of MLOps

## Model-centric view

Collect what data you can, and develop a model good enough to deal with the noise in the data.

Hold the data fixed and iteratively improve the code/model.

## Data-centric view

The consistency of the data is paramount. Use tools to improve the data quality; this will allow multiple models to do well.

*Hold the code fixed and iteratively improve the data.*

***MLOps' most important task is to make high quality data available through all stages of the ML project lifecycle.***

1. [MLOps: From Model-centric to Data-centric AI](#)
2. [MLOps: Andrew Ng](#)

# How to start?

- Learn python / R
- Revised your maths (probability and linear algebra) and statistics knowledge
- Take a few good data science / ML / DL course
  - Learn EDA (Exploratory Data Analysis)
  - Basic feature engineering
  - Linear regression to neural networks with underlying maths
  - Try to implement the algorithm by yourself
- Basic SQL
- Write whatever you learn

# How to start?

- Choose an use case / one area you like
- Search for a dataset (maybe a tabular data)
- Identity the problem to solve
- Identity if it is a classification or regression problem
- Do some EDA to understand the data
- Some statistical analysis to understand the data
- Identify the inputs and output - dependent and independent variables
- Take a basic model - linear regression or logistic regression
- Try to implement the algorithm by yourself - loss function - optimization algorithm
- Now import that model from Scikit-learn and check the accuracy (or other measure)
- Try to optimize your code

Participate in the  
different data science  
competitions

# Kaggle Report | State of Data Science 2021

[Click here for the report](#)

# Useful links

- Monitoring the model training & dataset versioning
  - <https://wandb.ai/site>
- Visualization of ML techniques :
  - egfycat (<https://gifycat.com/gifs/search/gradient+descent>)
- Courses: (Paid)
  - [Foundations of data science, One Fourth Labs, IITM Research Park](#)
  - [IBM Data Science Course](#)
  - [Deep Learning, One Fourth Labs, IITM Research Park](#)
  - [Andrew Ng: Deep learning specializations](#)
- Courses in YouTube:
  - [Andrew Ng: Machine Learning](#)
  - [Mithesh Khapra, IITM: Deep Learning](#)
  - [NYU Deep Learning Spring 2021: Yann LeCun, Alfredo Canziani](#)
- [Awesome ML courses](#)
- <https://www.deeplearning.ai/>
- [By Microsoft](#) | [By Amazon](#)
- [ML for security](#)
- <https://scikit-learn.org/>

# Useful links

- CNN materials: cs231n.stanford.edu (<http://cs231n.stanford.edu/>)
- Andrew Ng DL notes: cs230.stanford.edu (<https://cs230.stanford.edu>)
- Machine Learning Yearning (<https://d2wvfoqc9gyqzf.cloudfront.net/content/uploads/2018/09/NgMLY01-13.pdf>)
- [CS224S: Spoken Language Processing: Stanford University](#)
- [Speech and Language Processing: Stanford University](#)
- [CS 329S: Machine Learning Systems Design: Stanford University](#)
- [Probability4datascience](#)
- Medium articles for data science
- <https://towardsdatascience.com/>
- <https://www.analyticsvidhya.com/>
- Jay Alammar: <https://jalammar.github.io/>
- <https://distill.pub>
- [Brief DL book from meta scientists:](#)
- [Free ebooks:](#)

## Twitter handle (or thread) / facebook / discord server / slack channel to follow

- <https://twitter.com/AndrewYNg> | <https://twitter.com/ylecun> | <https://twitter.com/drfeifei>
- Great for ML/DL resources (**Alfredo Canziani**): <https://twitter.com/alfcnz>
- Lex Fridman (his podcasts are excellent): <https://twitter.com/lexfridman>
- <https://twitter.com/abhi1thakur>
- MIT Tech Review: <https://twitter.com/techreview>
- <https://twitter.com/TDataScience>
- <https://twitter.com/AnalyticsVidhya>
- <https://roundtable.datascience.salon/top-data-science-machine-learning-slack-communities>
- <https://towardsdatascience.com/top-20-data-science-discord-servers-to-join-in-2020-567b45738e9d>
- <https://discuss.huggingface.co/t/join-the-hugging-face-discord/11263>
- <https://twitter.com/mlopscommunity> | [https://twitter.com/ml\\_india](https://twitter.com/ml_india)
- <https://www.facebook.com/groups/machinelearningforum>
- <https://www.facebook.com/groups/1294016480653992>
- <https://www.reddit.com/r/MachineLearning/>



# Useful links

- dbworld (<https://research.cs.wisc.edu/dbworld/>)
- ml-news (<https://groups.google.com/forum/#!categories/ml-news>)
- ir-list (<https://sigir.org/sig-irlist/>)
- <https://groups.google.com/g/naacl-latin-america>

# TOP DATA SCIENCE & AI TRENDS FOR 2022



**1. MLOPS WILL ACQUIRE WEIGHT IN OPERATIONALISING AND SCALING AI/ML**

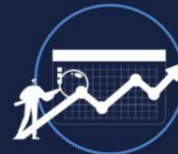


**6. HOLISTIC DATA FABRICS WILL BE AT THE CENTRE OF REDEFINING DATA STRATEGY**

**2. RESPONSIBLE AI WILL GAIN SIGNIFICANCE IN INDIA**



**7. DATA SCIENCE INDUSTRY WILL SEE A CONSOLIDATION OF MORE IPOs OR ACQUISITIONS**



**3. DEARTH OF DATA ENGINEERS WILL BE FELT MORE THAN DATA SCIENTISTS**



**8. ANALYTICS PROFESSIONALS WILL COMMAND HIGHER SALARIES**

**4. LARGE LANGUAGE MODELS WILL BECOME LARGER**



**9. NEW AGE ALGORITHMS WILL SEE HIGHER UTILISATION IN INDUSTRIES**



**5. FURTHER FORMALISATION OF DATA SCIENCE EDUCATION WILL LEAD TO SPECIALISED COURSES**



**10. EFFORTS IN AI/ML LOCALISATION WILL IMPROVE**

## **Graduate courses**

### **IIM Ahmedabad**

IIM Ahmedabad launched the Brij Disa Centre for Data Science and Artificial Intelligence (CDSA). It said that this centre would conduct research in data science and artificial intelligence that will contribute to businesses, governance, and policymaking. It will also publish an exhaustive annual report on the data science and AI industry in India that will look at the challenges that exist in the industry and offer solutions to overcome them.

### **IIM Nagpur**

IIM Nagpur launched the Post Graduate Certificate Programme in Data Science for Business Excellence and Innovation. It is aimed towards early and mid-career professionals seeking upskilling opportunities to perform effectively in their careers. It ranges from 9 to 12 months in duration and is scheduled beyond work hours and with flexibility of learning from convenient locations from working professionals.

### **IIT-I and IIM-I**

IIT Indore and IIM Indore have signed an MoU to offer a two-year Master of Science programme in Data Science and Management. The course is targeted at working executives and fresh graduates. The two institutes said that the aim of the course is to equip students with the right skills needed for data scientists roles like data management skills, project management, big data analytics life cycle, and systems thinking.

### IIT Guwahati

IIT Guwahati has launched a Bachelor's course in Data Science and Artificial Intelligence. It said that the first batch of 20 students would be admitted to the institution through the JEE Advanced-2021 counselling process. It is designed to train students holistically, focusing on relevant courses from other disciplines like computer science, electronics and electrical engineering, Mathematics and Statistics, the institute said.

### IIT Delhi

The School of Artificial Intelligence (ScAI) at IIT Delhi announced a post-graduate programme in Artificial Intelligence named 'M.Tech in Machine Intelligence & Data Science (MINDS)'. It is expected to begin in July 2022. M.Tech in MINDS is an industry-sponsored programme and the focus will be to work on industry-relevant AI problems. They will be co-advised by an IIT Delhi professor and a researcher from the sponsoring company. The eligibility to pursue this programme will be an undergraduate degree in science or engineering.

### IIT Patna

It has introduced three new courses under the four-year B Tech program for the academic session in early 2021, one of which is for Artificial Intelligence and Data Science. Admission to these courses will be through JEE Advanced 2021. It will

### IIT Ropar

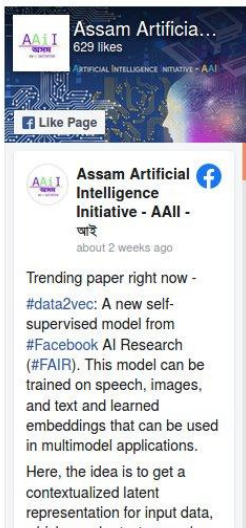
The Punjab Skill Development Mission and IIT Ropar have teamed up to offer a free course in artificial intelligence and data science. It will be held online with IIT professors and industry leaders participating in it. It is open for Class 12th students with a mathematics background. Students will get a certificate from IIT Ropar after course completion



## Join us at AAIL

Exciting opportunities to work and volunteer with us at **AAIL**. We have immediate openings for the volunteering roles and discipline on a variety of areas, such as **Event moderation**, **Content creation**, **Graphic designing**, **A.I. project execution** and many more. For more details, refer [here](#).

## Assam AI Initiative - AAIL - আই



### Latest Events

Data Science during the Covid-19 Pandemic with emphasis on Healthcare

## Assam Artificial Intelligence Initiative

Assam Artificial Intelligence Initiative (AAIL – আই) is a non-profit initiative to promote various scopes of AI and build a network for students/researchers/industry persons to exchange their knowledge. We will also try to popularize AI among school children/non-specialists of this region by organizing popular talks and demonstrating different AI applications through this initiative. This is a platform to promote AI for the betterment of our society.





Questions?

You can reach me @

- [Linkedin](#)