# Data Science using Python

Manaranjan Pradhan

A Data Enthusiast

LinkedIn: https://www.linkedin.com/in/manaranjanpradhan
He writes blogs at www.awesomestats.in

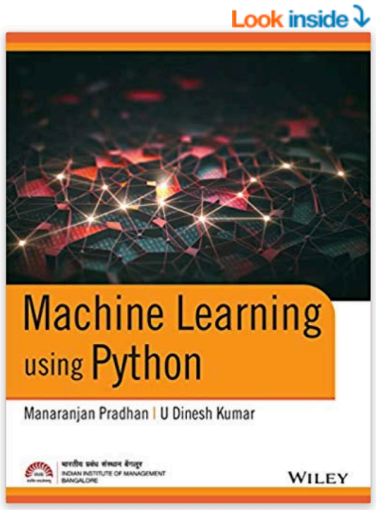
# Complete the following steps…

- Create a working directory on your desktop / laptop, where you can store all data and the programs of this workshop
- Download datasets from the link below and store in the above directory
  - https://github.com/manaranjanp/**ML_Python**
  - Download the complete repository as zip file and unarchive it
- Install latest Anaconda Distribution for Python (3.6+) on your desktop / laptop
  - https://www.anaconda.com/download/

# About Me

- Worked for HP and iGate for 12+ years
- Working as freelancer for last 5 years
- Mostly conduct workshops on Spark, Databricks, Machine Learning and Deep Learning
- https://www.linkedin.com/in/manaranjanpradhan

**Manaranjan Pradhan (Manu)**




**Machine Learning using Python** Paperback – 2019

by U Dinesh Kumar Manaranjan Pradhan (Author)

★★★★☆ 7 customer reviews

› See all 2 formats and editions

Kindle Edition ₹ 423.20 — Read with Our Free App

Paperback ₹ 529.00 prime — 2 New from ₹ 529.00

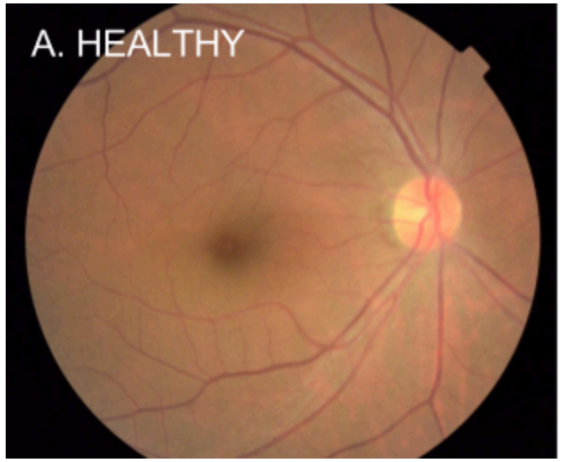
This book is written to provide a strong foundation in machine learning using Python libraries by providing real-life case studies and examples. It covers topics such as foundations of machine learning, introduction to Python, descriptive analytics and predictive analytics. Advanced machine learning concepts such as decision tree learning, random forest, boosting, recommended systems, and text analytics are covered. The book takes a balanced approach between theoretical understanding and practical applications. All the topics include real-world examples and provide step-by-step approach on how to explore, build, evaluate, and optimize machine learning models.

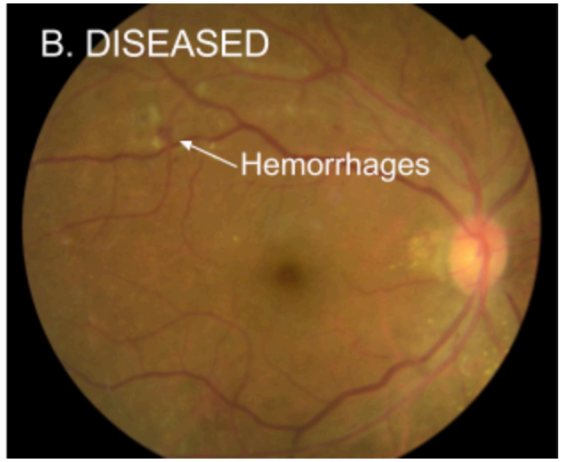
https://www.amazon.in/Machine-Learning-Python-Manaranjan-Pradhan-ebook/dp/B07RLQPNRX/

# Course Outline

| | |
|---|---|
| Introduction to Data Science and Setting up data analysis environment | Introduction to Data Science<br>Setting up Python Environment for Data Analysis<br><br>Overview of Data Analysis Stack - Numpy, Pandas, Matplotlib, scipy and Scikit-learn |
| Accessing and preparing data with Pandas | Loading data from Different Sources<br>Data manipulation - Filtering, Grouping, Ordering, Joining<br>Dealing with missing Data |
| Data Exploration, Visualizations & Statistical Analysis | Histograms, Bar charts<br>Density Plots, Box Plots, Scatter Plots, Heat Maps<br>Understanding Basic Statistics, Distributions, Correlations |
| Algorithms for Regression and Classification Problems | Understanding loss function and gradient descent approach for loss minimization<br>Linear Regression, Logistic Regression, Decision Trees, KNN<br>Model Optimization & Parameter Tuning |
| Clustering | K-means clustering<br>Finding optimal number of clusters |
| Model Evaluation | Creating Training, validation and Test Data Sets<br>Cross validations<br>Understanding Evaluation Metrics: RMSE, R-square, ROC, Confusion Matrix, Precision, Recall, Accuracy etc. |

# Detection of Diabetic Eye Disease




**haemorrhage**
/ˈhɛmərɪdʒ/

*noun*
plural noun: **hemorrhages**

1. an escape of blood from a ruptured blood vessel.
   "a massive haemorrhage of the brain"

https://research.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html

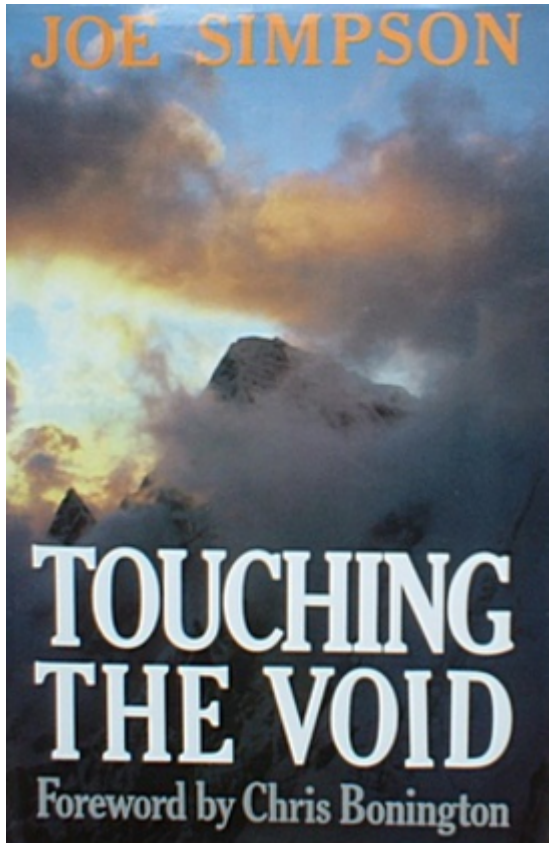
- Google working closely with doctors both in India and the US, created a development dataset of 128,000 images which were each evaluated by 3-7 ophthalmologists from a panel of 54 ophthalmologists.
- Trained a deep neural network to detect referable diabetic retinopathy.
- Then tested the algorithm's performance on two separate clinical validation sets totaling ~12,000 images, with the majority decision of a panel 7 or 8 U.S. board-certified ophthalmologists serving as the reference standard.
- the algorithm has a F-score (combined sensitivity and specificity metric, with max=1) of 0.95, which is slightly better than the median F-score of the 8 ophthalmologists we consulted (measured at 0.91).

# Into Thin Air

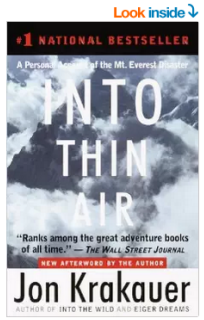

It details the author's presence at Mount Everest during the 1996 Mount Everest disaster, when eight climbers were killed and several others were stranded by a "rogue storm"

# Touching the Void

***Touching the Void*** is a 1988 book by Joe Simpson, recounting his and Simon Yates' successful but disastrous and nearly fatal climb of the 6,344-metre (20,813 foot) Siula Grande in the PeruvianAndes in 1985.



# Customers who bought this item also bought…




**Into Thin Air: A Personal Account of the Mt. Everest Disaster**

Paperback – October 19, 1999

by Jon Krakauer (Author, Photographer), Randy Rackliff (Illustrator), Daniel Rembert (Contributor), & 2 more

2,414 customer reviews

#1 Best Seller in Mountain Climbing

See all 65 formats and editions

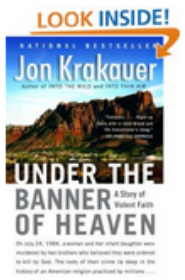
| Kindle | Hardcover | Paperback | Mass Market Paperback |
|---|---|---|---|
| $3.40 | $18.66 | $10.36 | from $0.01 |
| Read with our free app | 667 Used from $0.01<br>82 New from $4.49<br>51 Collectible from $6.37 | 483 Used from $0.01<br>117 New from $6.22<br>11 Collectible from $9.70 | 487 Used from $0.01<br>18 New from $4.59<br>19 Collectible from $3.00 |

**National Bestseller**

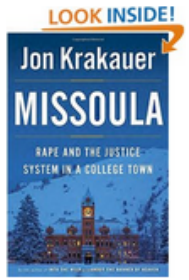
A bank of clouds was assembling on the not-so-distant horizon, but journalist-mountaineer Jon

## Customers Who Bought This Item Also Bought

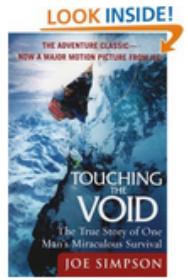

Into the Wild
› Jon Krakauer
2,304
#1 Best Seller in Travelogues & Travel Essays
Paperback
$7.34 Prime



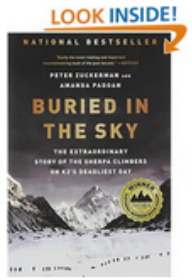

Under the Banner of…
› Jon Krakauer
1,361
Paperback
$10.03 Prime
Get it by Tomorrow




Missoula: Rape and the…
Jon Krakauer
361
Hardcover
$18.09 Prime
Get it by Tomorrow




Touching the Void: The…
› Joe Simpson
315
Paperback
$11.22 Prime
Get it by Tomorrow




Buried in the Sky: The…
› Peter Zuckerman
225
Paperback
$10.63 Prime
Get it by Tomorrow

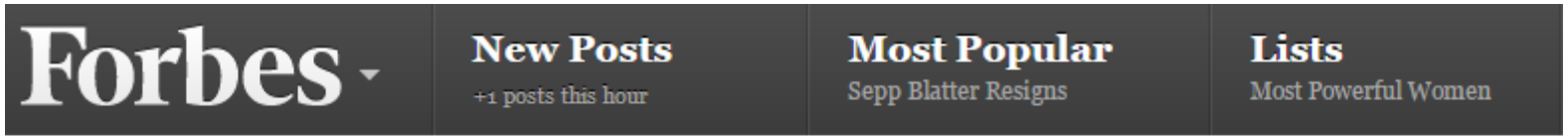
# Recommendations are key to personalization

## Amazon's recommendation secret

McKinsey estimated that 35 percent of consumer purchases on Amazon come from product recommendations, although the e-commerce giant itself has never revealed its own estimates. In 2016, it offered its open-source artificial intelligence (AI) framework called, DSSTNE (pronounced as "destiny"), for free to encourage the development of artificial intelligence apps.

https://martechtoday.com/roi-recommendation-engines-marketing-205787

# What customers bought together




4/06/1998 @ 12:00AM

Share

## Diaper-beer syndrome

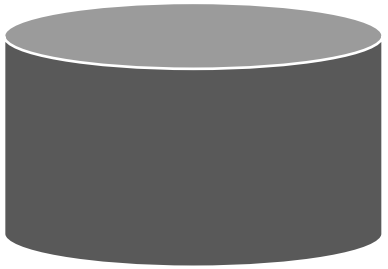
IT'S PART OF the folklore of data processing. A retail chain put all its checkout-counter data into a giant digital warehouse and set the disk drives spinning.

Out popped a most unexpected correlation: sales of diapers and beer.

Evidently, young fathers would make a late-night run to the store to pick up Pampers and get some Bud Light while they were there.

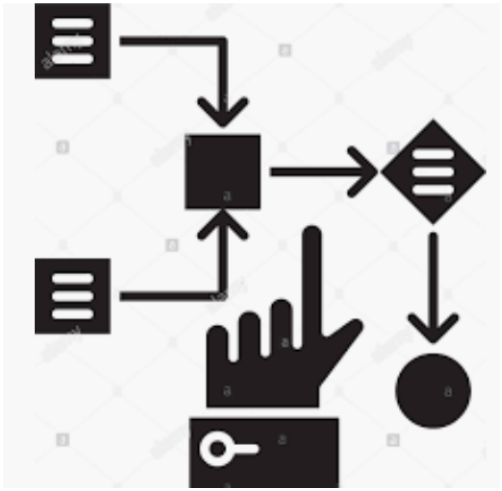
Capitalizing on the discovery, the store placed the disparate items together. Sales zoomed.

# Key Components



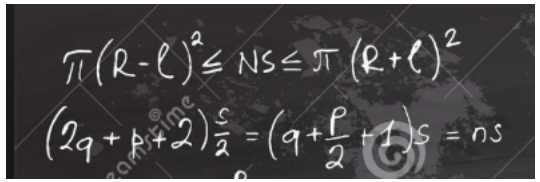
**Data**

- Historical Evidences
- Samples representing problem scenario



**Learning Algorithm**

- Statistical Learning
- Machine Learning or Deep Learning



**Model**

A mathematical expression of the pattern or evidence found in data and can be used to find insights and applied in future to predict.

# How AI, ML and DL are related?

## Artificial Intelligence

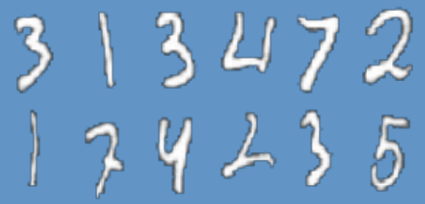
Any technique that enables computers to mimic human behavior

## Machine Learning

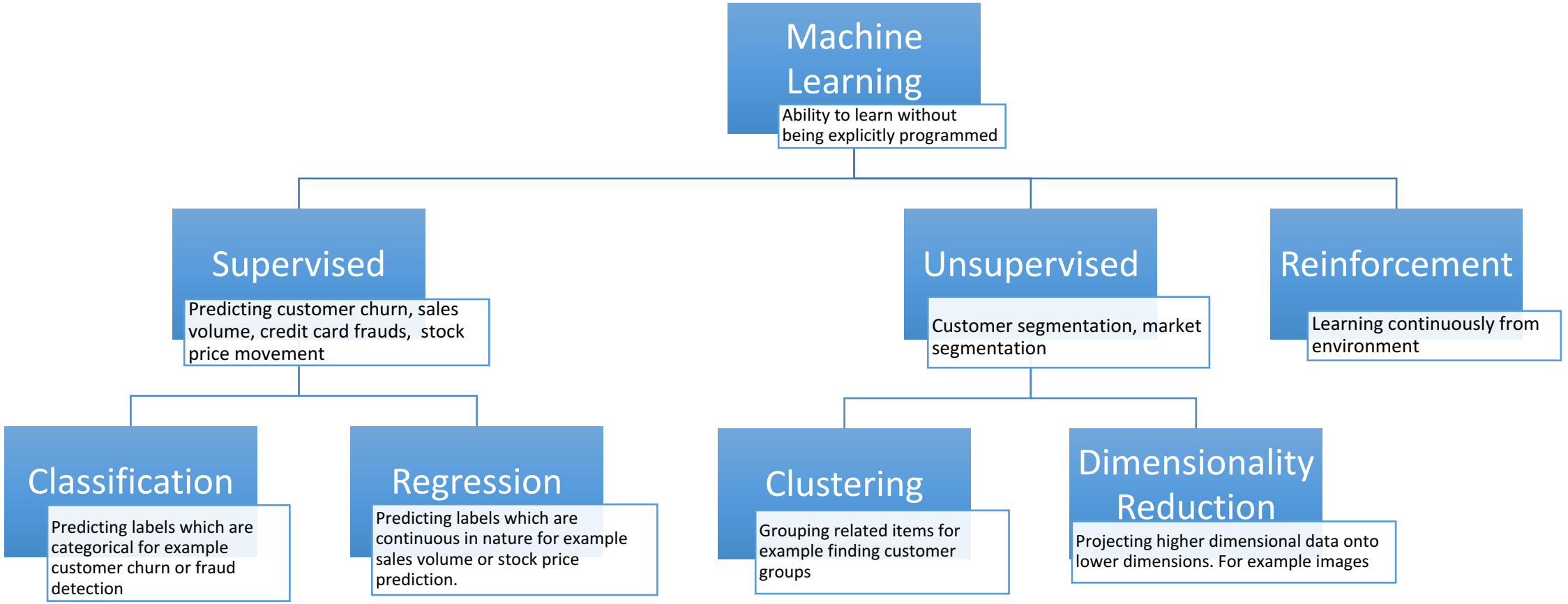
Ability to learn without explicitly being programmed

## Deep Learning

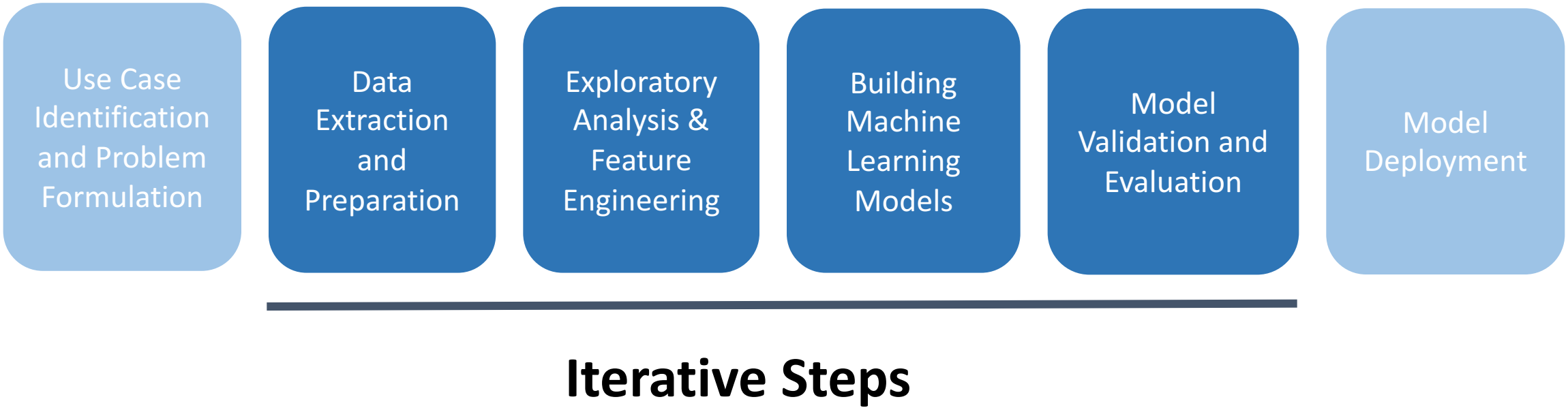
Learn underlying features in data using neural networks



# Machine Learning Algorithms

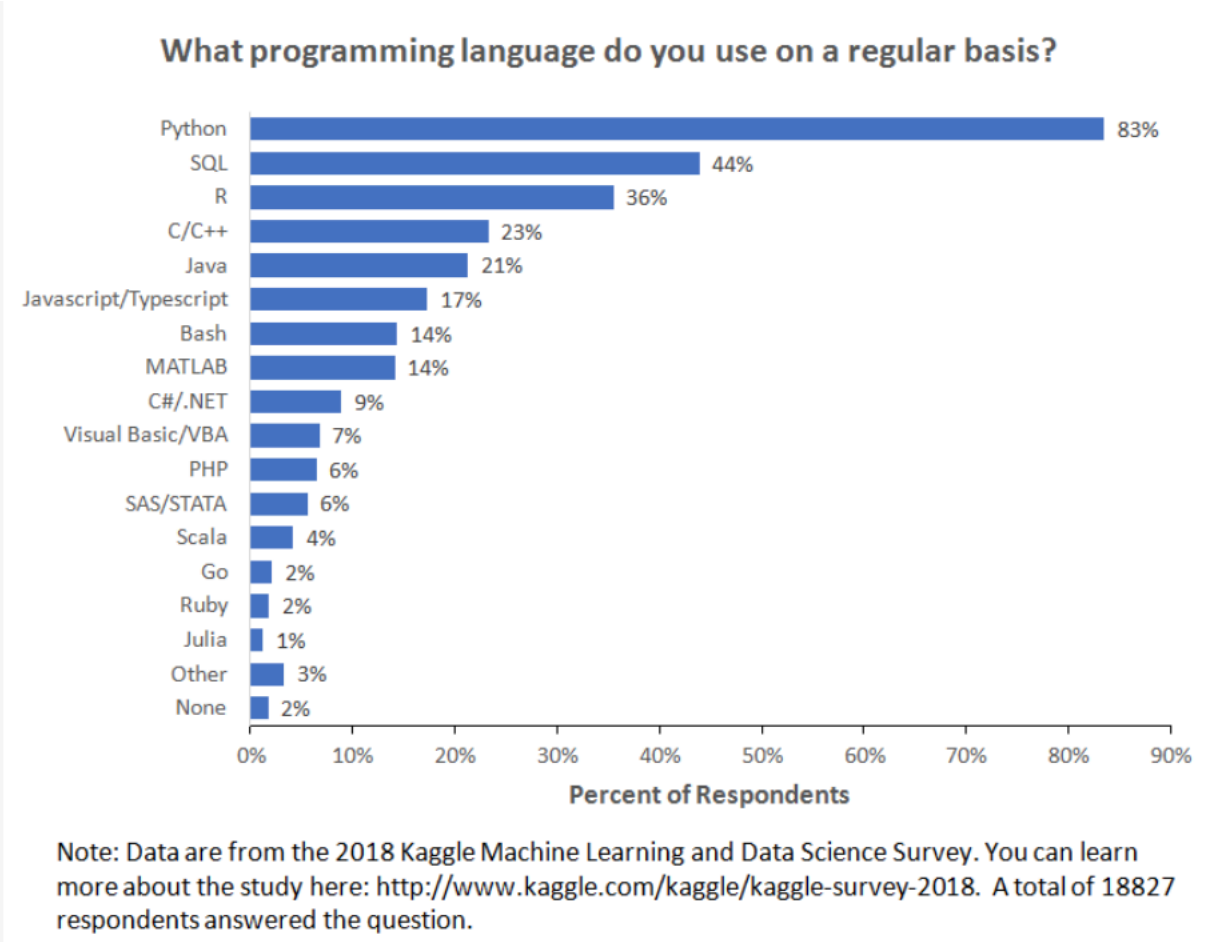


# Lifecycle


Use Case Identification and Problem Formulation
Data Extraction and Preparation
Exploratory Analysis & Feature Engineering
Building Machine Learning Models
Model Validation and Evaluation
Model Deployment
Iterative Steps

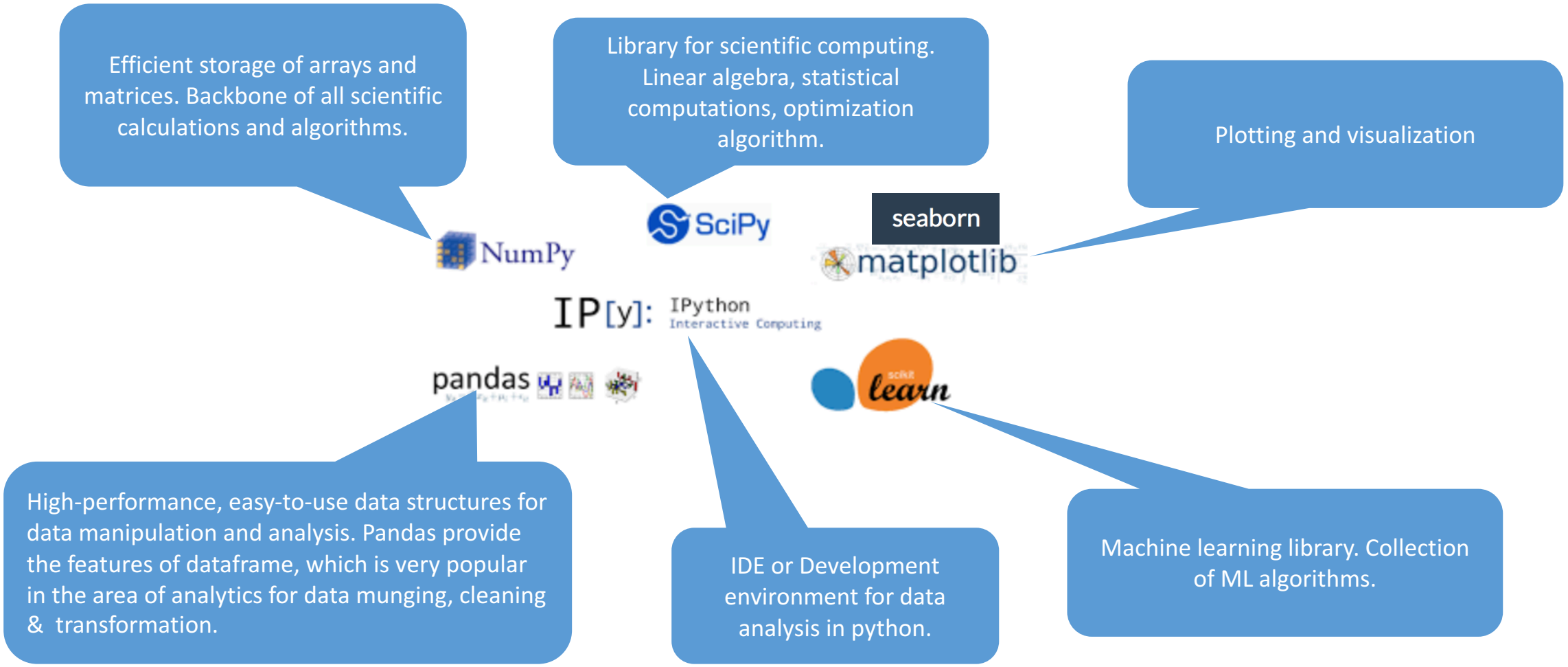
# What Tools are available?

# Language for Machine Learning

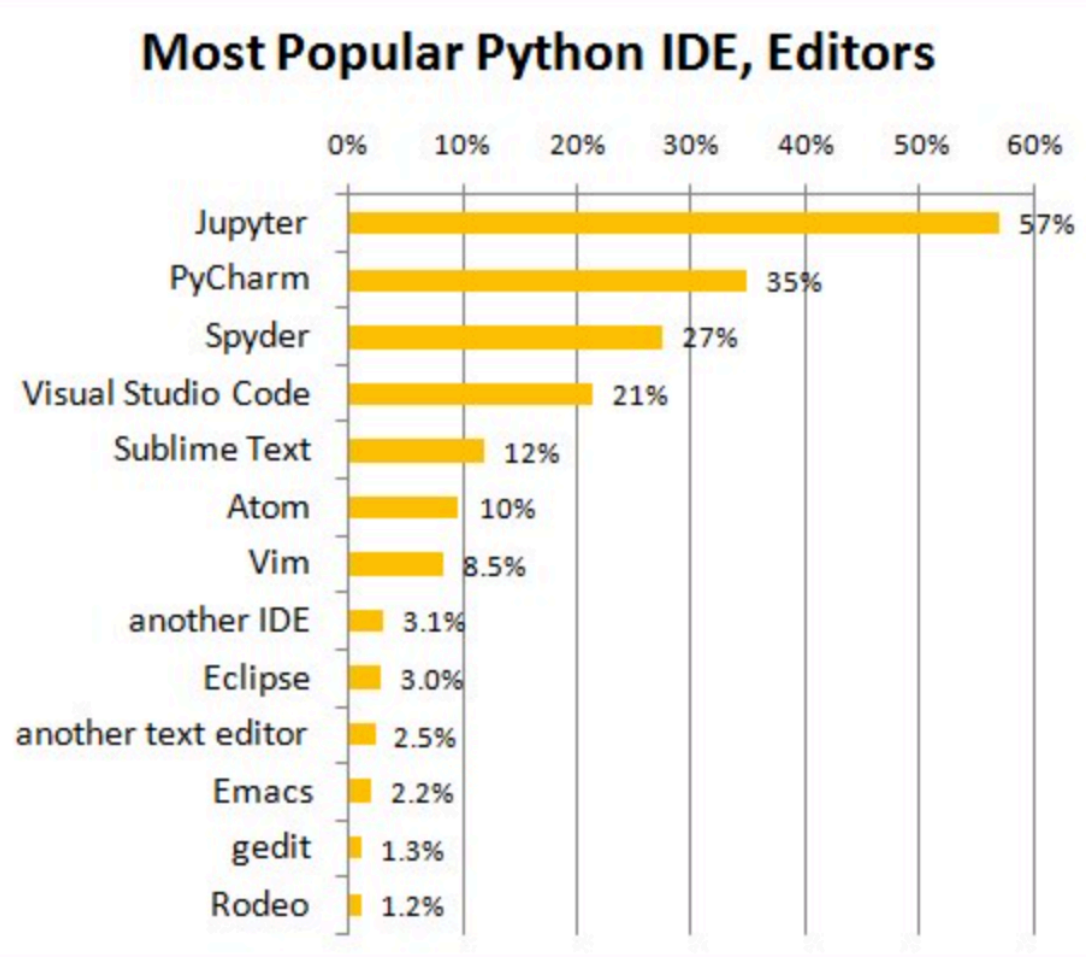



https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/

# Python Stack For Data Science

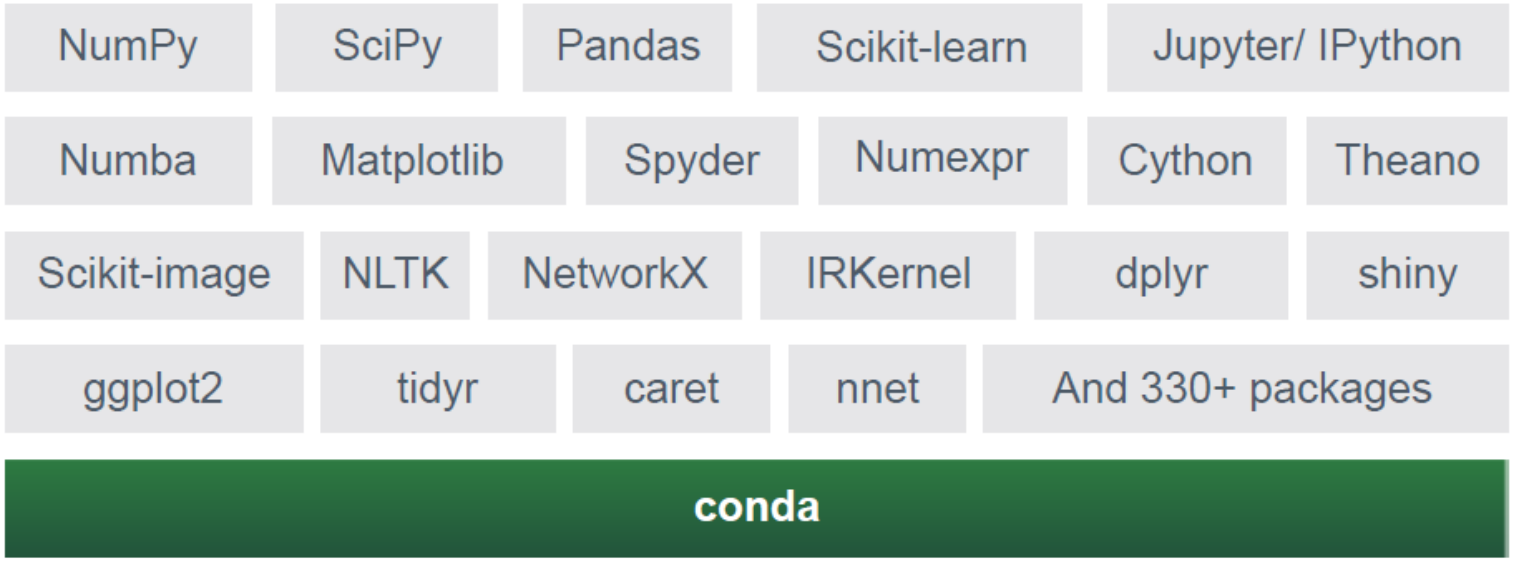


# Language for Machine Learning




https://www.anaconda.com/download/

# Python Distribution




**Game-Changing Enterprise Ready Python Distribution**

- 2 million downloads in last 2 years
- 200k / month and growing
- conda package manager serves up 5 *million* packages per month
- Recommended installer for IPython/Jupyter, Pandas, SciPy, Scikit-learn, etc.

Source: Continuum Analytics

[Download link: https://www.continuum.io/downloads](https://www.continuum.io/downloads)

# Start Jupyter notebook

- For MAC
  - Click on Anaconda Navigator and click on “launch notebook”
  - Or go to command prompt and enter
    - **jupyter notebook**
- For Windows
  - Go to anaconda command prompt and enter
    - **jupyter notebook**

# Start a jupyter notebook


**Click on new to start  new notebook. For every hands on exercise, start a new notebook.**