DATA SCIENCE FROM A RESEARCH PERSPECTIVE

Dr.Vani Vasudevan, Professor – CSE,

Nitte Meenakshi Institute of Technology, Bangalore.

Some Quotes!!!

- The purpose of computing is **insight**, not numbers.– Richard W. Hamming (**Data Science**)
- A data scientist is someone who knows more **statistics** than a computer scientist and more **computer science** than a statistician.– Josh Blumenstock (**Mathematics**)
- On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" . . . I am not able to rightly apprehend the kind of confusion of ideas that could provoke such a question.– Charles Babbage (**Data Wrangling**)
- Money is a **scoreboard** where you can rank how you're doing against other people.– Mark Cuban (**Measures**)
- It is easy to lie with statistics, but easier to lie without them. (**Statistical Analysis**)
- At their best, **graphics** are instruments for reasoning.– Edward Tufte (**Data Visualization**)

Some More Quotes!!!

- All models are wrong, but some **models are useful**.– George Box (**Mathematical Models**)
- Any sufficiently advanced form of cheating is indistinguishable from **learning**.– Jan Schaumann (**Machine Learning**)
- A change in **quantity** also entails a **change in quality**.– Friedrich Engel (**Big Data**)
  https://www.internetlivestats.com/

Data Science...

Data science is an emerging field that

(1) is extremely **transdisciplinary** —bridging between the theoretical, computational, experimental, and biosocial areas;

(2) deals with enormous **amounts** of complex, incongruent, and dynamic **data** from multiple sources; and

Dr. Vani Vasudevan

4

11/12/22

Source : Data Science and Predictive Analytics
Ivo D. Dinov, "Biomedical and Health Applications using R", Springer 2018

Data Science

(3) aims to develop algorithms, methods, tools, and services capable of ingesting such datasets and generating semiautomated decision support systems.

The latter can mine the data for patterns or motifs, predict expected outcomes, suggest clustering or labeling of retrospective or prospective observations, compute data signatures or fingerprints, extract valuable information, and offer evidence-based actionable knowledge.

Data science techniques often involve data manipulation (wrangling), data harmonization and aggregation, exploratory or confirmatory data analyses, predictive analytics, validation, and fine-tuning.

Source : Data Science and Predictive Analytics
Ivo D. Dinov, “Biomedical and Health Applications using R”, Springer 2018

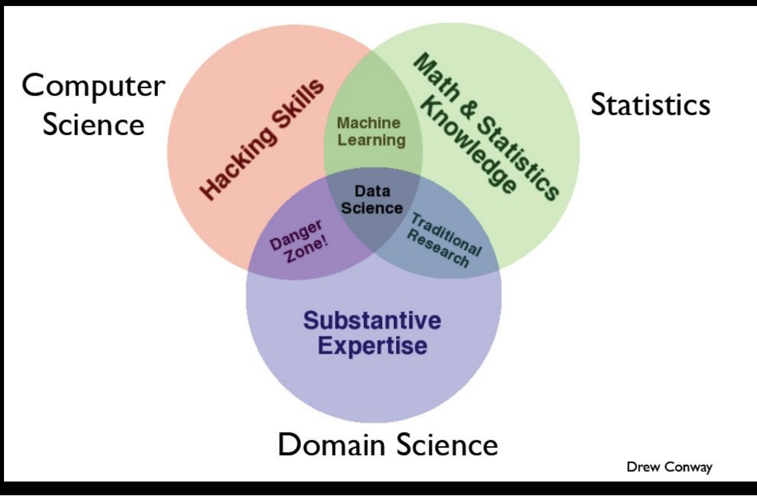
What is Data Science?

Like any emerging field, it isn't yet well defined, but incorporates elements of:

- Exploratory Data Analysis and Visualization
- Machine Learning and Statistics
- High-Performance Computing technologies for dealing with scale.

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

SKILL SETS FOR DATA SCIENCE



source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.
Some more resources:
https://towardsdatascience.com/introduction-to-statistics-e9d72d818745
https://deepai.org/machine-learning-glossary-and-terms/data-science
https://www.devopsschool.com/blog/what-is-data-science-advantages-and-disadvantages-of-data-science/

Appreciating Data

- Computer Scientists do not naturally appreciate data: it's just stuff to run through a program.
- The usual way to test algorithm performance is to run the implementation on "random data".
- But, interesting data sets are a scarce resource, which requires hard work and imagination to obtain.

Dr. Vani Vasudevan

8

11/12/22

source: Steven S. Skiena, "The Data Science Design Manual", Springer 2017.

Computer Vs. Real Scientists

- Scientists strive to understand the complicated and messy natural world, while **computer scientists build their own clean and organized virtual worlds.** Thus:
- Nothing is ever completely true or false in science, while everything is either true or false in Computer Science / Mathematics.
- Scientists are data-driven, while **computer scientists are algorithm-driven**.
- Scientists obsess about discovering things, which **computer scientists invent rather than discover.**
- Scientists are comfortable with the idea that data has errors; **computer scientists are not**.

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Genius Vs. Wisdom

- Software developers are hired to produce code.
- Data Scientists are hired to produce insights.
- Genius shows in finding the right answer!!!
- Wisdom shows in avoiding the wrong answers.

Data science (like most things) benefits more from wisdom than from genius.

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Developing Wisdom

- Wisdom comes from **experience**.
- Wisdom comes from **general knowledge**.
- Wisdom comes from **listening to others**.
- Wisdom comes from **humility**, observing how often you have been **wrong and why/how**.

I seek pass on wisdom, through my experience on the difficulty of making good predictions.

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Developing Curiosity

- The good **data scientist** develops a curiosity about the domain/application they are working in.
- They **talk shop with the people whose data** they are working on.
- They **read the newspaper** every day, to get a broader perspective on the world.

Dr. Vani Vasudevan

12

11/12/22

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Asking GOOD QUESTIONS:

Software developers are not encouraged to ask questions, but data scientists are:

- What exciting things might you be able to learn from a given data set?
- What things do you/your people really want to know?
- What data sets might get you there?

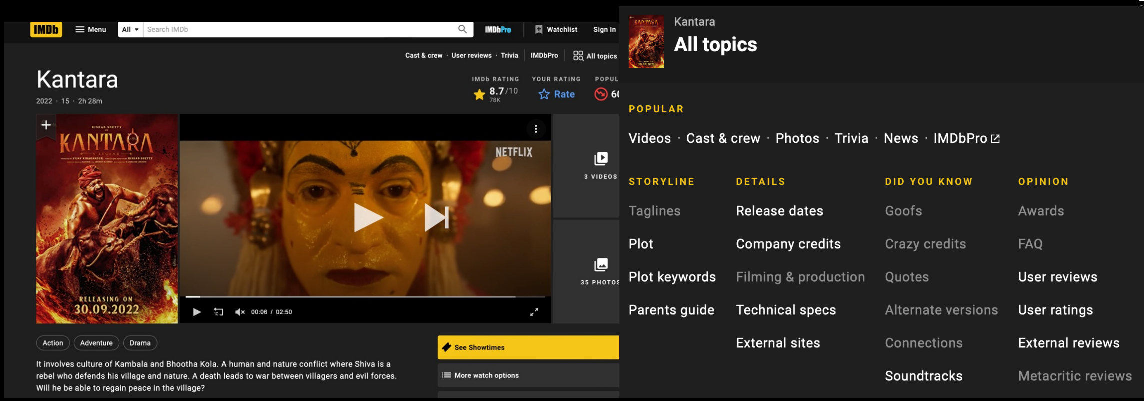
source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

LET'S PRACTICE ASKING QUESTIONS!

- Who, What, Where, When, and Why on the following datasets:

1. International Movie Database (IMDB)
2. New York City Taxi Trip Duration

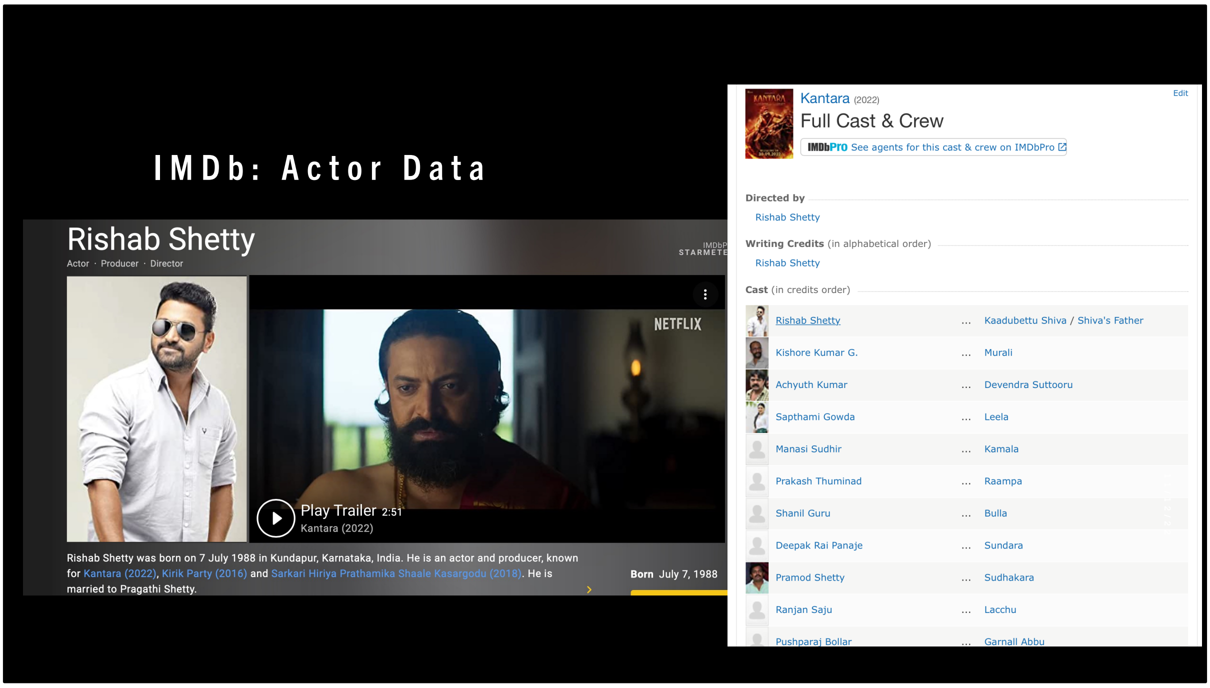
IMDb: Movie Data



Dr. Vani Vasud

15

11/12/22


IMDb: Actor Data
Rishab Shetty
Actor · Producer · Director
NETFLIX
Play Trailer 2:51
Kantara (2022)
Rishab Shetty was born on 7 July 1988 in Kundapur, Karnataka, India. He is an actor and producer, known for Kantara (2022), Kirik Party (2016) and Sarkari Hiriya Prathamika Shaale Kasargodu (2018). He is married to Pragathi Shetty.
Born July 7, 1988
Kantara (2022)
Full Cast & Crew
IMDbPro See agents for this cast & crew on IMDbPro
Edit
Directed by
Rishab Shetty
Writing Credits (in alphabetical order)
Rishab Shetty
Cast (in credits order)
Rishab Shetty ... Kaadubettu Shiva / Shiva's Father
Kishore Kumar G. ... Murali
Achyuth Kumar ... Devendra Suttooru
Sapthami Gowda ... Leela
Manasi Sudhir ... Kamala
Prakash Thuminad ... Raampa
Shanil Guru ... Bulla
Deepak Rai Panaje ... Sundara
Pramod Shetty ... Sudhakara
Ranjan Saju ... Lacchu
Pushparaj Bollar ... Garnall Abbu

Movie Questions

- Can we predict how well people will like a movie? What about its gross?
- What does the social network of actors look like?
- What is the age distribution of actors and actresses in film?
- Do stars live longer or shorter lives than the bit players or public?

Dr. Vani Vasudevan

17

11/12/22

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

NYC Taxi Trip Data

- https://www.kaggle.com/competitions/nyc-taxi-trip-duration/data

Data fields

- **id** - a unique identifier for each trip
- **vendor_id** - a code indicating the provider associated with the trip record
- **pickup_datetime** - date and time when the meter was engaged
- **dropoff_datetime** - date and time when the meter was disengaged
- **passenger_count** - the number of passengers in the vehicle (driver entered value)
- **pickup_longitude** - the longitude where the meter was engaged
- **pickup_latitude** - the latitude where the meter was engaged
- **dropoff_longitude** - the longitude where the meter was disengaged
- **dropoff_latitude** - the latitude where the meter was disengaged
- **store_and_fwd_flag** - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- **trip_duration** - duration of the trip in seconds

NYC Taxi Trip Questions

- https://www.kaggle.com/competitions/nyc-taxi-trip-duration/data
- How far do they travel?
- How much slower is traffic during rush hour?
- Where are people traveling to/from at different times of the day?
- Where should drivers go to pick up their next fare?

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Properties Of Data

- Structured vs. Unstructured Data
- Quantitative vs. Categorical Data
- Big Data vs. Little Data

- Do not blindly aspire to analyze large data sets. Seek the right data to answer a given question, not necessarily the biggest thing you can get your hands on.

Dr. Vani Vasudevan

20

11/12/22

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Classification And Regression

- Two types of problems arise repeatedly in **traditional data science and pattern recognition applications**, the challenges of classification and regression.

Classification: Often, we seek to assign a label to an item from a discrete set of possibilities. Such problems as predicting the winner of a particular sporting contest (team A or team B?) or deciding the genre of a given movie (comedy, drama, or animation?) are classification problems, since each entail selecting a label from the possible choices.

Classification And Regression

- Two types of problems arise repeatedly in **traditional data science and pattern recognition applications**, the challenges of classification and regression.

Regression: Another common task is to forecast a given numerical quantity. Predicting a person's weight or how much rain we will get this year is a regression problem, where we forecast the future value of a numerical function in terms of previous values and other relevant features.

Dr. Vani Vasudevan

22

11/12/22

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Classification And Regression

- The best way to see the intended distinction is to look at a variety of data science problems and label (classify) them as regression or classification.
- Different algorithmic methods are used to solve these two types of problems.

1. Will the price of a particular stock be higher or lower tomorrow?
2. What will the price of a particular stock be tomorrow?
3. Is this person a good risk to sell an insurance policy to?
4. How long do we expect this person to live?

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

PRACTICE QUESTION

Identifying Data Sets

1. Identify where interesting data sets relevant to the following domains can be found on the web:

(a) Books.

(c) Stock prices.

(d) Risks of diseases.

(e) Colleges and universities.

(f) Crime rates.

For each of these data sources, explain what you must do to turn this data intoa usable format on your computer for analysis.

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

The Data Science Pipeline

1. Get or collect data
2. Manipulate and process data
3. Modeling and analysis
4. Visualize, evaluate, present, and communicate

Dr. Vani Vasudevan

11/12/22

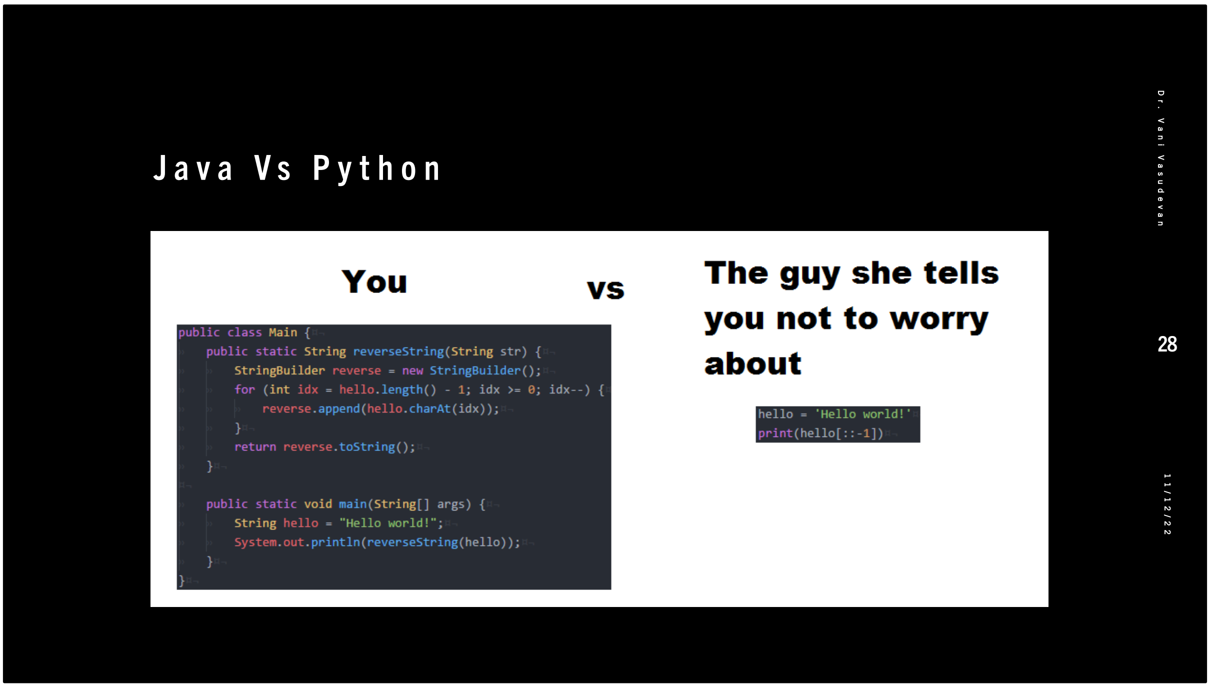
Some Useful Web Resources to Kick Start Your Learning And Research!

- https://cognitiveclass.ai/ - Data Science and Cognitive Computing Courses
- https://www.kdnuggets.com/ - Site on AI, Analytics, Big Data, Data Mining, Data Science, and ML https://www.kaggle.com/ - ML & DS community
- https://data.gov/ - US government data
- http://archive.ics.uci.edu/ml/index.php - ML Repository
- https://homepages.ecs.vuw.ac.nz/~marslast/MLbook.html - Stephen Marsland homepage
- https://www.cs.waikato.ac.nz/ml/weka/courses.html – Waikato University - Weka MOOC
- https://nptel.ac.in/courses/106/106/106106202/ - NPTEL - Machine Learning
- Rohit singh, tommi jaakkola, and ali mohammad. *6.867 Machine learning.* Fall 2006. Massachusetts institute of technology: MIT opencourseware, https://ocw.mit.edu.
- Leslie kaelbling, tomás lozano-pérez, isaac chuang, and duane boning. *6.036 introduction to machine learning.* Fall 2020. Massachusetts institute of technology: MIT opencourseware, https://ocw.mit.edu.
- https://ocw.mit.edu/courses/hst-953-collaborative-data-science-for-healthcare-fall-2020/ collaborative data science for healthcare
- https://ocw.mit.edu/courses/15-062-data-mining-spring-2003/

Data Science Tools

1. Python(Most known)

- Python is one of the most dominant languages in the field of data science today because of its flexibility, ease of use in terms of syntax, open-source nature, and ability to handle, clean, manipulate, visualize, and analyze data.
- Python was essentially developed as a programming language. However, it offers a wide range of libraries, such as TensorFlow, Keras, PyTorch, Seaborn, etc., that are attractive for both programmers and data scientists alike. Moreover, there are various other tools connected to and built with the help of Python, such as Dask, SciPy, Cython, Matplotlib, and High-Performance Analytics Toolkit(HPAT).



source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Python Libraries

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

The Python Ecosystem

- **Numpy**: N-dimensional arrays, Matrices and Linear Algebra
- **Scipy**: Algorithms from linear algebra, optimization, statistics and signal processing
- **Pandas**: Data Manipulation and Analysis
- **Matplotlib**: Data Visualization
- **IPython**: Interactive shell for Python
- **Scikit-learn**: Machine Learning

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Anaconda

- A bundle of data science, machine learning and visualization libraries.
- Contains every library you'd need in this course.
- Easiest way to avoid inter-dependency issues.
- Installation

1. Go to: https://www.anaconda.com/distribution/
2. Download the installer for your OS and Python version of choice and follow instructions

Dr. Vani Vasudevan

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

Jupiter Notebook

- A browser-based notebook with support for code, text, mathematical expressions, inline plots and other rich media

COLAB

- Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud.
- With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.

Dr. Vani Vasudevan

33

11/12/22

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.

PANDAS

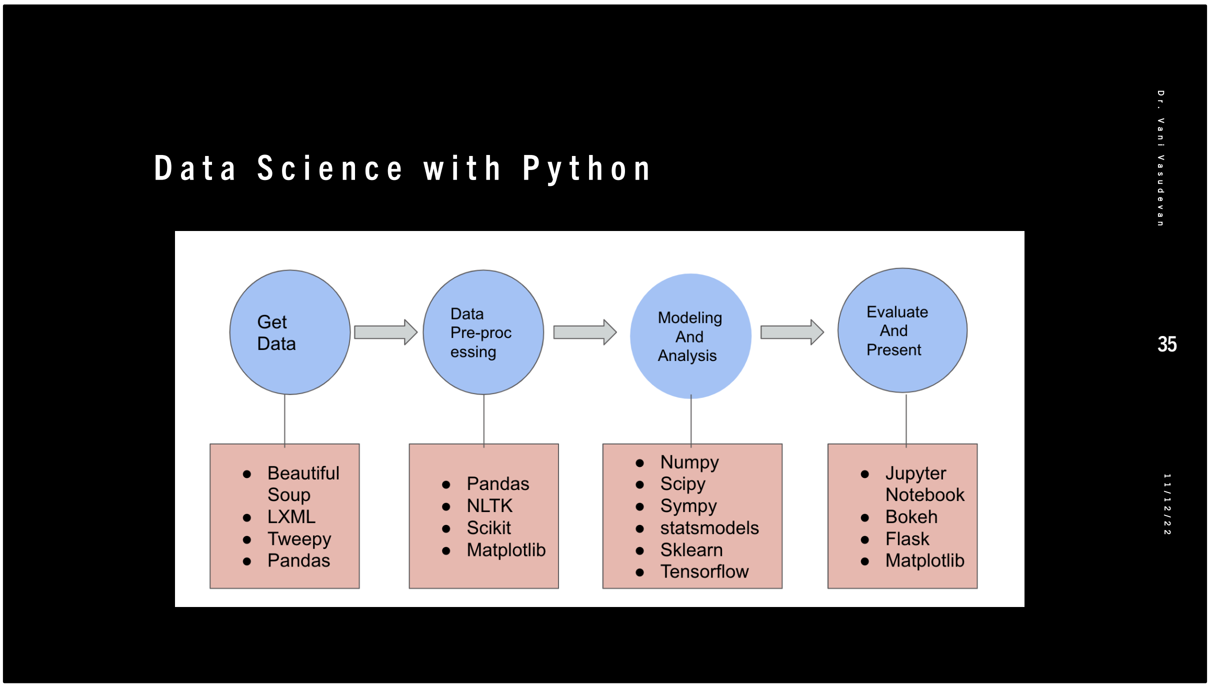
- It is a library for data manipulation and analysis
- Data structures: Series and Data Frame (tabular data)
- Data is loaded in-memory, hence super fast (but not ideal for datasets of scale > memory size)

Dr. Vani Vasudevan

34

11/12/22

source: Steven S. Skiena, “The Data Science Design Manual”, Springer 2017.



1.source: Steven S. Skiena, "The Data Science Design Manual", Springer 2017.
2.Data Accquisition --- Beautiful Soup, LXML, Scrapy, Tweepy (Obtaining data by spidering the web etc), pySpark (for large data), mySQL client, mongoDB
3.Pre-processing -- Domain Specific Pre-processing techniques (Text: NLTK, Images: Scikit-image etc)
4.Analysis/Modeling

1. Exploratory Data Analysis: Pandas
2. Visualization: pylab, matplotlib, seaborn
3. Modeling: numpy, scipy, sympy
4. Hypothesis Testing: scipy, statsmodels
5. Machine Learning: sklearn --

5.Evaluation/Interpretation/Communication

1. Latex in Ipython
2. Bokeh
3. Flask

Other Data Science Tools

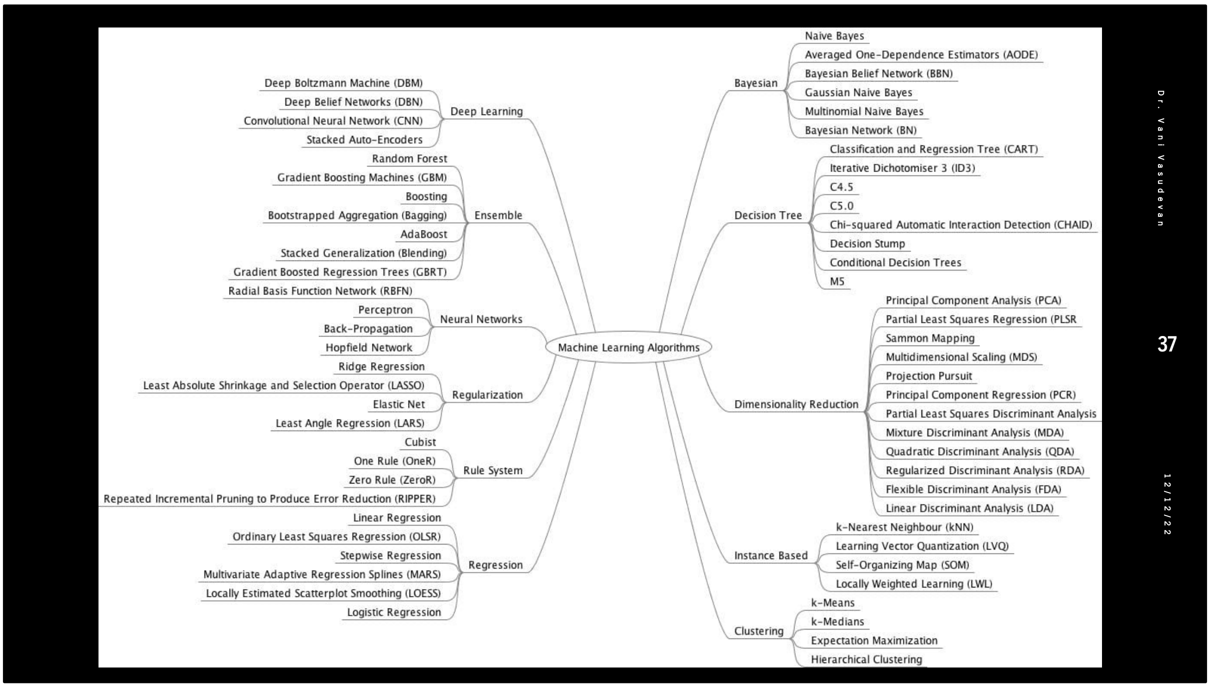
- WEKA
- R (RStudio)
- MATLAB
- Statistical Analysis System (SAS)
- Apache Hadoop
- Tableau
- QlikView
- RapidMiner
- Excel
- PowerBI
- Google Analytics

and much more!

Dr. Vani Vasudevan

36

11/12/22



Source: Machine Learning Mastery: https://machinelearningmastery.com/wp-content/uploads/2021/03/MachineLearningAlgorithms.jpg?__s=hbkixgpvleicleslspeo&utm_source=drip&utm_medium=email&utm_campaign=MMLA+Mini-Course&utm_content=Machine+Learning+Algorithms+Mind-Map+and+Mini-Course

https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/

Parametric Approaches :

Logistic Regression
Linear Discriminant Analysis
Perceptron
Naive Bayes
Simple Neural Networks

Non Parametric Approaches:
k-Nearest Neighbors
Decision Trees like CART and C4.5

Support Vector Machines

RESEARCH PROBLEMS IN DATA SCIENCE AND BIG DATA

The research problems in intersection of big data with data science

- Approaches to make the models learn with a smaller number of data samples
- Neural Machine Translation to Local languages
- Handling Data and Model drift for real-world applications
- Handling interpretability of deep learning models in real-time applications
- Building large scale generative based conversational systems
- Building context-sensitive large-scale systems

https://towardsdatascience.com/top-20-latest-research-problems-in-big-data-and-data-science-c6fb51e03136

RESEARCH PROBLEMS IN DATA SCIENCE AND BIG DATA

The research problems related to data engineering aspects

- Lightweight Big Data analytics as a Service
- Auto conversion of algorithms to MapReduce problems

https://towardsdatascience.com/top-20-latest-research-problems-in-big-data-and-data-science-c6fb51e03136

RESEARCH PROBLEMS IN DATA SCIENCE AND BIG DATA

The problems related to core big data area of handling the scale:

- Scalable architectures for parallel data processing
- Handling real-time video analytics in a distributed cloud
- Efficient graph processing at scale

https://towardsdatascience.com/top-20-latest-research-problems-in-big-data-and-data-science-c6fb51e03136

RESEARCH PROBLEMS IN DATA SCIENCE AND BIG DATA

The research problems to handle noise and uncertainty in the data:

- Identify fake news in near real-time
- Dimensional Reduction approaches for large scale data
- Training / Inference in noisy environments and incomplete data
- Handling uncertainty in big data processing

Dr. Vani Vasudevan

41

12/12/22

https://towardsdatascience.com/top-20-latest-research-problems-in-big-data-and-data-science-c6fb51e03136

RESEARCH PROBLEMS IN DATA SCIENCE AND BIG DATA

The research problems in the security and privacy area:

- Anomaly Detection in Very Large-Scale Systems
- Effective anonymization of sensitive fields in the large-scale systems
- Secure federated learning with real-world applications
- Scalable privacy preservation on big data

Dr. Vani Vasudevan

12/12/22

https://towardsdatascience.com/top-20-latest-research-problems-in-big-data-and-data-science-c6fb51e03136

Ten Research Challenge Areas In Data Science

1. Scientific Understanding of Learning, Especially Deep Learning Algorithms.
2. Causal Reasoning
3. Precious Data
4. Multiple, Heterogeneous Data Sources
5. Inferring From Noisy and/or Incomplete Data
6. Trustworthy AI
7. Computing Systems for Data-Intensive Applications
8. Automating Front-End Stages of the Data Life Cycle
9. Privacy
10. Ethics

Dr. Vani Vasudevan

43

11/12/22

https://hdsr.mitpress.mit.edu/pub/d9j96ne4/release/3

ANY QUESTIONS?

Reach me @ vani.v@nmit.ac.in
LinkedIn: https://in.linkedin.com/in/dr-vani-vasudevan-0b89b713