data science glossary

Apache
- RDD
- DataFrame
- Dataset
- MLlib
- ML Pipelines
- GraphX
- Spark Streaming
- Structured Streaming
- spark-packages.org
- Catalyst Optimizer
- Tungsten
- Continuous Applications
- In-memory computing

Hadoop
- MapReduce
- Hadoop Distributed File System (HDFS)
- Yet Another Resource Negotiator (YARN)
- HBase
- Hive
- Apache Pig
- Apache Spark
- Sqoop
- Oozie
- ZooKeeper
- Apache Flume
- Hue
- Mahout
- Ambari
- Hadoop Common

Cloud Computing
- XaaS (Anything-as-a-Service)
- Software-as-a-Service (SaaS)
- Platform-as-a-Service (PaaS)
- Infrastructure-as-a-Service (IaaS)
- Public Cloud
- Private Cloud
- Hybrid Cloud
- AWS
- Amazon EC2 (Elastic Cloud Compute)
- Amazon Simple Storage Service (S3)
- Cloud Sourcing
- Consumer Cloud
- Multi-tenancy
- Vertical Cloud
- Cloud Portability
- Cloud Backup
- Cloud Enablement
- Cloud Migration
- Cloudstorming
- Cloud Broker

Descriptive Statistics
- Population
- Sample
- Parameter
- Statistic
- Generalizability
- Distribution
- Mean
- Median
- Mode
- Skew
- Range
- Variance
- Standard Deviation
- Interquartile Range (IQR)

deep learning
- Artificial Neural Networks (ANNs)
- Biological Neuron
- Perceptron
- Multilayer Perceptron (MLP)
- Feedforward Neural Network
- Recurrent Neural Network
- Activation Function
- Backpropagation
- Cost Function
- Gradient Descent
- Vanishing Gradient Problem
- Convolutional Neural Network
- Long Short Term Memory Network (LSTM)

Machine Learning
- Classification
- Regression
- Clustering
- Association
- Decision Trees
- Support Vector Machines
- Neural Networks
- Deep Learning
- Reinforcement Learning
- (k-fold) Cross-validation
- Bayesian
- Random Forest

Natural Language Processing
- Tokenization
- Normalization
- Stemming
- Lemmatization
- Corpus
- Stop Words
- Parts-of-speech (POS) Tagging
- Statistical Language Modeling
- Bag of Words
- n-grams
- Regular Expressions
- Zipf's Law
- Similarity Measures
- Syntactic Analysis
- Semantic Analysis
- Sentiment Analysis
- Information Retrieval

Internet of Things (IoT)
- 6LoWPAN
- Advanced Encryption Standard (AES)
- Application Programming Interface (API)
- Bluetooth Low Energy (BLE)
- Embedded Software
- Gateway
- General Packet Radio Service (GPRS)
- Industrial, Scientific, and Medical (ISM) Band
- Link Budget
- Machine to Machine (M2M)
- Media Access Control (MAC)

Predictive Analytics
- Predictive Model
- Artificial Intelligence
- Uplift Model
- Vast Search
- Automatic Suspect Discovery (ASD)

Database
- Relational Database
- Database Management System (DBMS)
- Primary Key
- Foreign Key
- Structured Query Language (SQL)
- NoSQL
- Metadata
- Consistency
- Data Redundancy
- ACID
- CAP Theorem
- Sharding
- Key-value Store
- Document Store
- Column-oriented Database
- Graph Database

Clustering
- Feature Selection
- Expectation Maximization (EM)
- Distance-based Methods
- Density- and Grid-Based Methods
- Matrix Factorization
- Spectral Methods
- Graph-based Techniques
- Streaming scenario

Big Data
- Big Data Volume
- Big Data Velocity
- Big Data Variety
- Big Data Veracity
- Big Data Variability
- Big Data Value
- Predictive Analytics
- Descriptive Analytics
- Prescriptive Analytics
- Database
- Data Warehouse
- ETL
- Business Intelligence
- Apache Hadoop
- Apache Spark
- Data lake
- Data mining
- Data preparation
- Data vault
- Data munging
- Data wrangling
- Data governance
- Data stewardship
- Data visualization
- Data Storytelling

data science glossary

1. Natural Language Processing

1.1. Tokenization

1.2. Normalization

1.3. Stemming

1.4. Lemmatization

1.5. Corpus

1.6. Stop Words

1.7. Parts-of-speech (POS) Tagging

1.8. Statistical Language Modeling

1.9. Bag of Words

1.10. n-grams

1.11. Regular Expressions

1.12. Zipf's Law

1.13. Similarity Measures

1.14. Syntactic Analysis

1.15. Semantic Analysis

1.16. Sentiment Analysis

1.17. Information Retrieval

2. Internet of Things (IoT)

2.1. 6LoWPAN

2.2. Advanced Encryption Standard (AES)

2.3. Application Programming Interface (API)

2.4. Bluetooth Low Energy (BLE)

2.5. Embedded Software

2.6. Gateway

2.7. General Packet Radio Service (GPRS)

2.8. Industrial, Scientific, and Medical (ISM) Band

2.9. Link Budget

2.10. Machine to Machine (M2M)

2.11. Media Access Control (MAC)

3. Predictive Analytics

3.1. Predictive Model

3.2. Artificial Intelligence

3.3. Uplift Model

3.4. Vast Search

3.5. Automatic Suspect Discovery (ASD)

4. Database

4.1. Relational Database

4.2. Database Management System (DBMS)

4.3. Primary Key

4.4. Foreign Key

4.5. Structured Query Language (SQL)

4.6. NoSQL

4.7. Metadata

4.8. Consistency

4.9. Data Redundancy

4.10. ACID

4.11. CAP Theorem

4.12. Sharding

4.13. Key-value Store

8.1. Artificial Neural Networks (ANNs)

8.2. Biological Neuron

8.3. Perceptron

8.4. Multilayer Perceptron (MLP)

8.5. Feedforward Neural Network

8.6. Recurrent Neural Network

8.7. Activation Function

8.8. Backpropagation

8.9. Cost Function

8.10. Gradient Descent

8.11. Vanishing Gradient Problem

8.12. Convolutional Neural Network

8.13. Long Short Term Memory Network (LSTM)

9. Descriptive Statistics

9.1. Population

9.2. Sample

9.3. Parameter

9.4. Statistic

9.5. Generalizability

9.6. Distribution

9.7. Mean

9.8. Median

9.9. Mode

9.10. Skew

9.11. Range

9.12. Variance

12.13. In-memory computing