Data Science

Data Intensive Computing

Dr. Akhter Raza

The Dawn of Big Data

- Google, Yahoo today
  - Web Search and Computational advertising
  - Google: 35,000 searches/sec
  - Yahoo! scale: 600 million users per month, 4 billion clicks per day, 25 terabytes of data collected every day
- Netflix 2007
  - Movie recommendations, netflix prize
  - 100 million ratings, 500,000 users, 18,000 movies
- Amazon 2003
  - Product recommendations, reviews
  - 29 million customers, millions of products
- Word Economic Forum 2011 at Davos
  - Personal data – digital data created by and about people – represents a new economic "asset class", touching all aspects of society

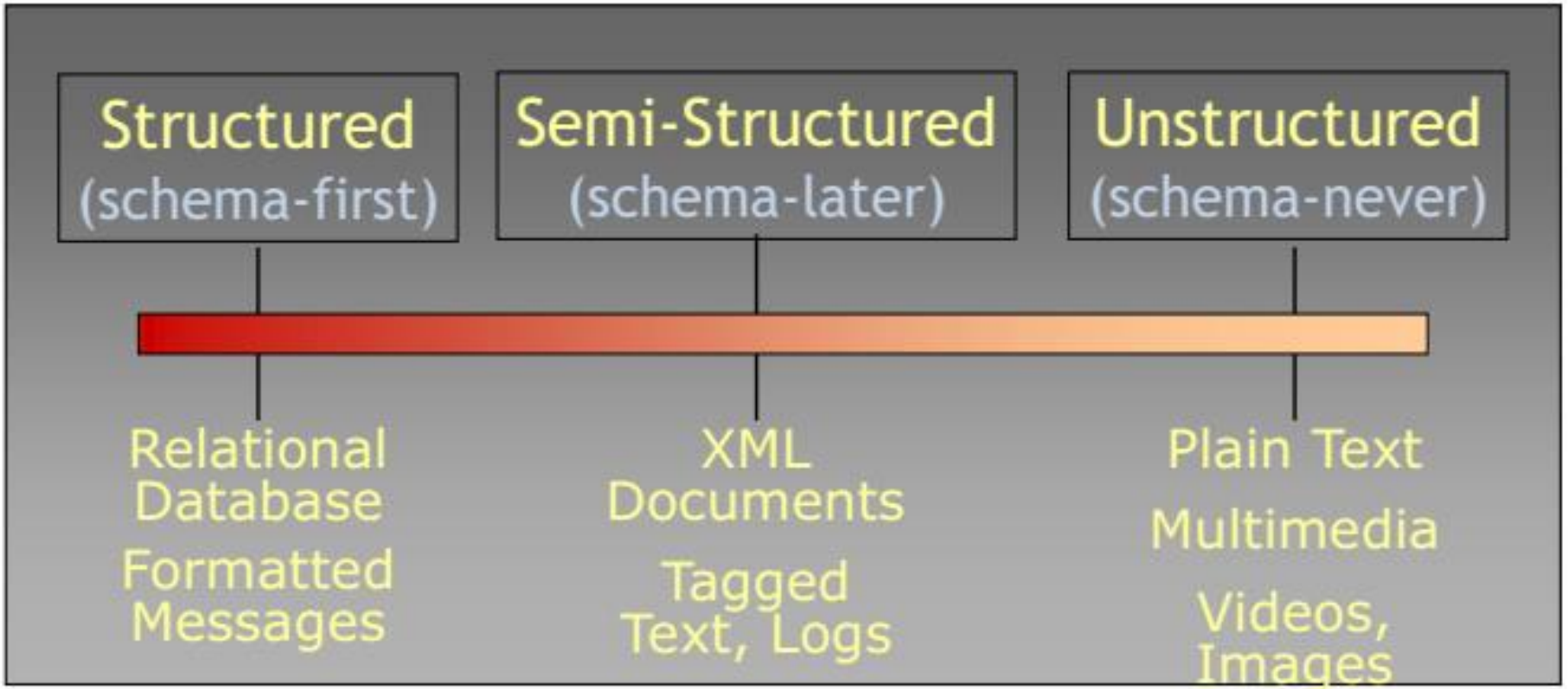
How Big is Your Data?

- Kilobyte (1000 bytes) 10^3
- Megabyte (1 000 000 bytes) 10^6
- Gigabyte (1 000 000 000 bytes) 10^9
- Terabyte (1 000 000 000 000 bytes) 10^{12}
- Petabyte (1 000 000 000 000 000 bytes) 10^{15}
- Exabyte (1 000 000 000 000 000 000 bytes) 10^{18}
- Zettabyte (1 000 000 000 000 000 000 000 bytes) 10^{21}
- Yottabyte (1 000 000 000 000 000 000 000 000 bytes) 10^{24}

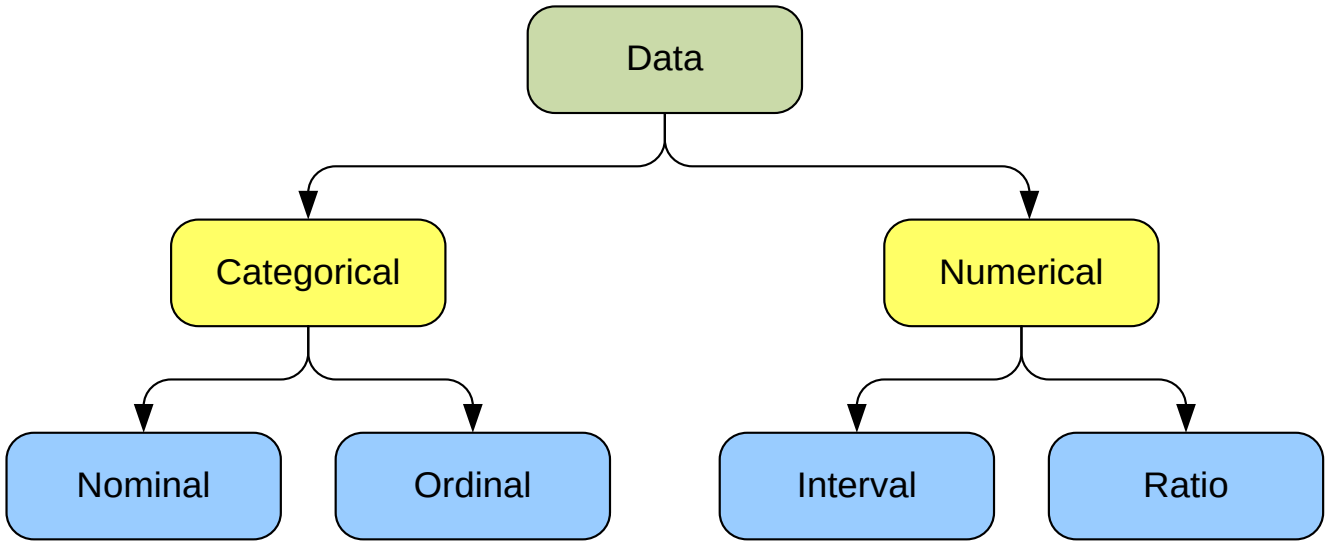
5 Vs of Big Data

- Raw Data: Volume
- Change over time: Velocity
- Data types: Variety
- Data Quality: Veracity
- Information for Decision Making: Value

The Structure Spectrum



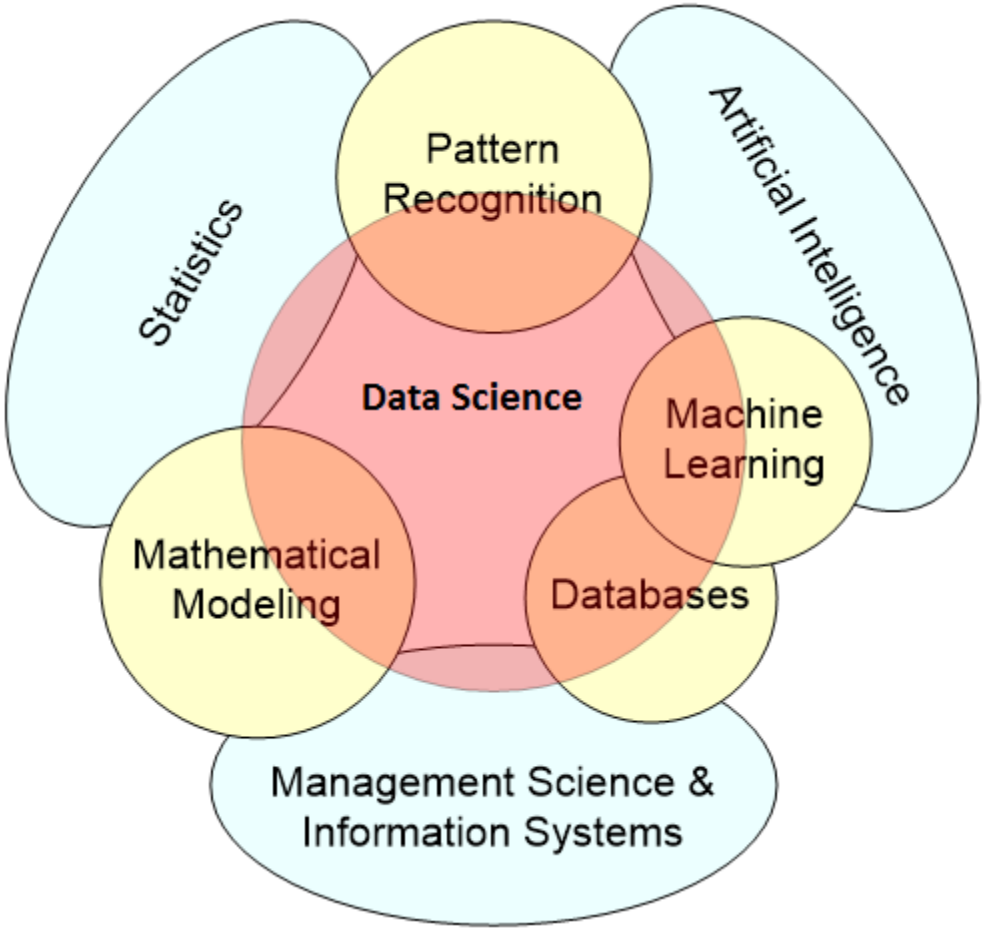
Data and its types



Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Data Science



Goal of Data Science

Turn data into data products.

Data to Data Products

Transaction Databases → Fraud Detection

Wireless Sensor Data → Smart Home

Social Media Data → Product Review and Consumer Satisfaction

Software Log Data → Automatic Trouble Shooting

Genotype and Phenotype Data → New treatment for Cancer

Top Applications

Loan Eligibility Prediction

Customer Segmentation

Traffic Sign Recognition

Image Caption Generator

Forest Fire Prediction

Road Lane Detection

Disease Prediction System

Driver Drowsiness Detection

Fake Job listing Detection

Other Data Products

Financial products for investment or retirement funds

Legal profession uses e-discovery tool for retrieval and review of legal documents

Political campaign management

Sports (e.g., Oakland baseball team)

Remote Sensing for Environment Monitoring

Data Products – Google

- Web Search
- Google Ads
- News Recommendation Engine
- Google Maps

Data Products – Netflix

- Personalized Movie Ratings
- Movie Recommendations
- Similar Movies
- Movie Categories (e.g., 80’s movie with a strong female lead, Kung Fu movies)

Data Products – LinkedIn/Facebook

- People you may know
- Applications you may like
- Jobs/Events you might be interested
- Classifier for bad users and bad content
- With high accuracy, Facebook can guess whether you are single or married

Who does not have LinkedIn/Facebook Account?

Data Products – Twitter

- Text Analysis – Spam Filter/Similarity Search
- User Sentiment/Satisfaction/Feedback
- News Breakout
- Trend and Topics

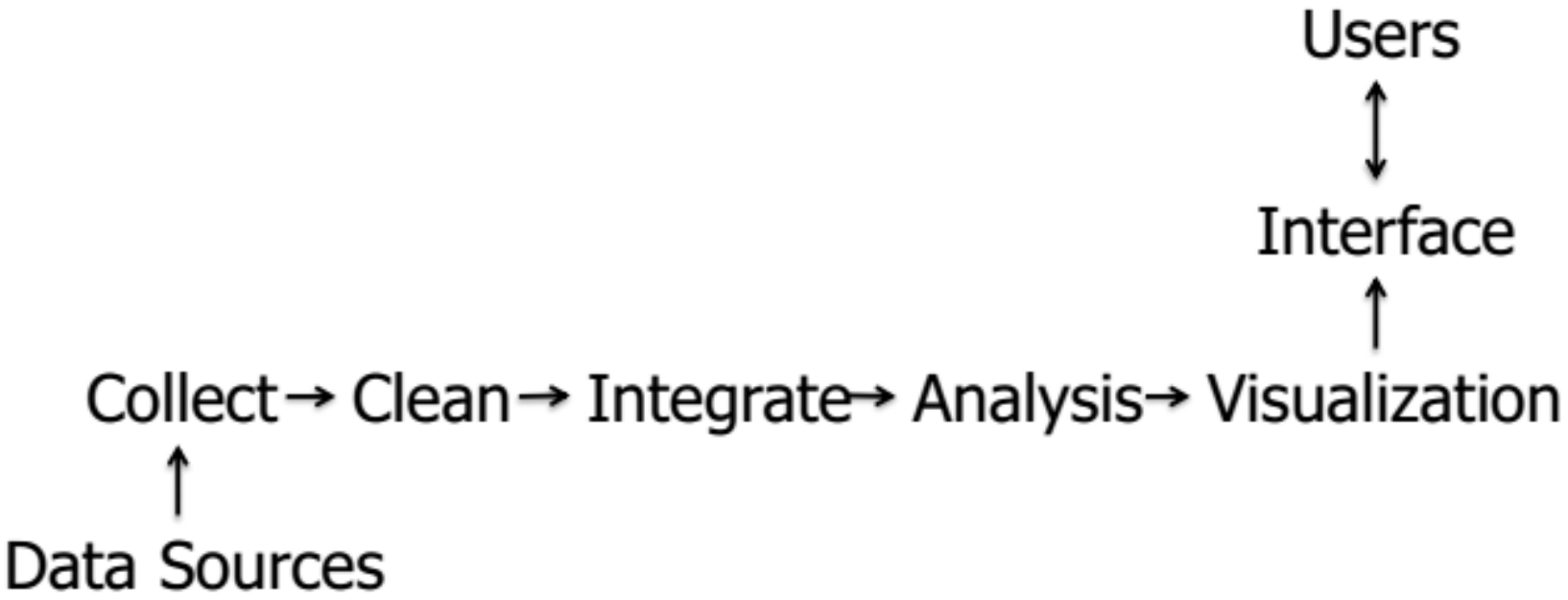
Data Products – Splunk

- Degradation, Failure Detection
- Identify Security Breach
- Event Monitoring
- Troubleshoot Tools
- Cross-platform Event Correlation

Splunk

- Grab data from many machines
- Index it
- Check for unusual events:
  - Disk problems
  - Network congestion
  - Security attacks
- Monitor Resources
  - Network
  - Memory usage
  - Disk use, latency
  - Threads
- Dashboard for cloud
  - administration.

The Life of Data (state-of-the-art)



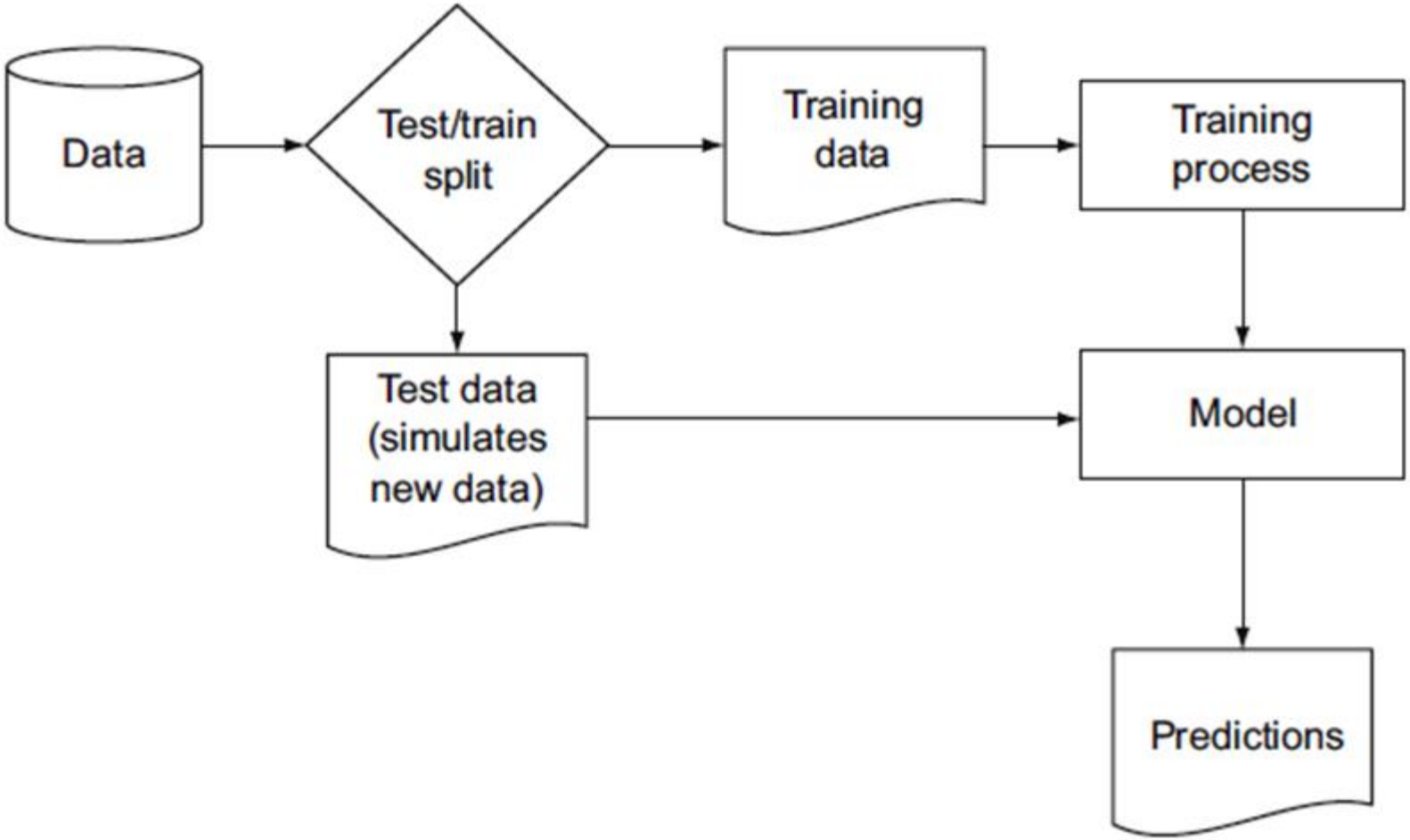
Challenges in Data Science

- Preparing Data (Noisy, Incomplete, Diverse, Streaming …)
- Analyze Data (Scalable, Accurate, Real time, Advanced Methods, Probabilities and Uncertainties ...)
- Represent Analysis Results (i.e. data product) (Story-telling, Interactive, explainable…)

Who's hiring Data Scientist?

- IT companies: Google, Twitter, Lexis/Nexis, Facebook, Pivotal/EMC…
- Media and Financial sectors Fox, CNN, NYT, Bloomburg,…
- Research: Biology, Medicine, Physics, Psychology,…
- Information office in government and corporations…
- Law firms: e-discovery tools…

Model Construction and Evaluation



Commonly used Models

Model	Description
Classification	Deciding if something belongs to one category or another
Scoring	Predicting or estimating a numeric value, such as a price or probability
Ranking	Learning to order items by preferences
Clustering	Grouping items into most-similar groups
Finding Relations	correlations or potential causes of effects seen in the data
Characterization	Very general plotting and report generation from data

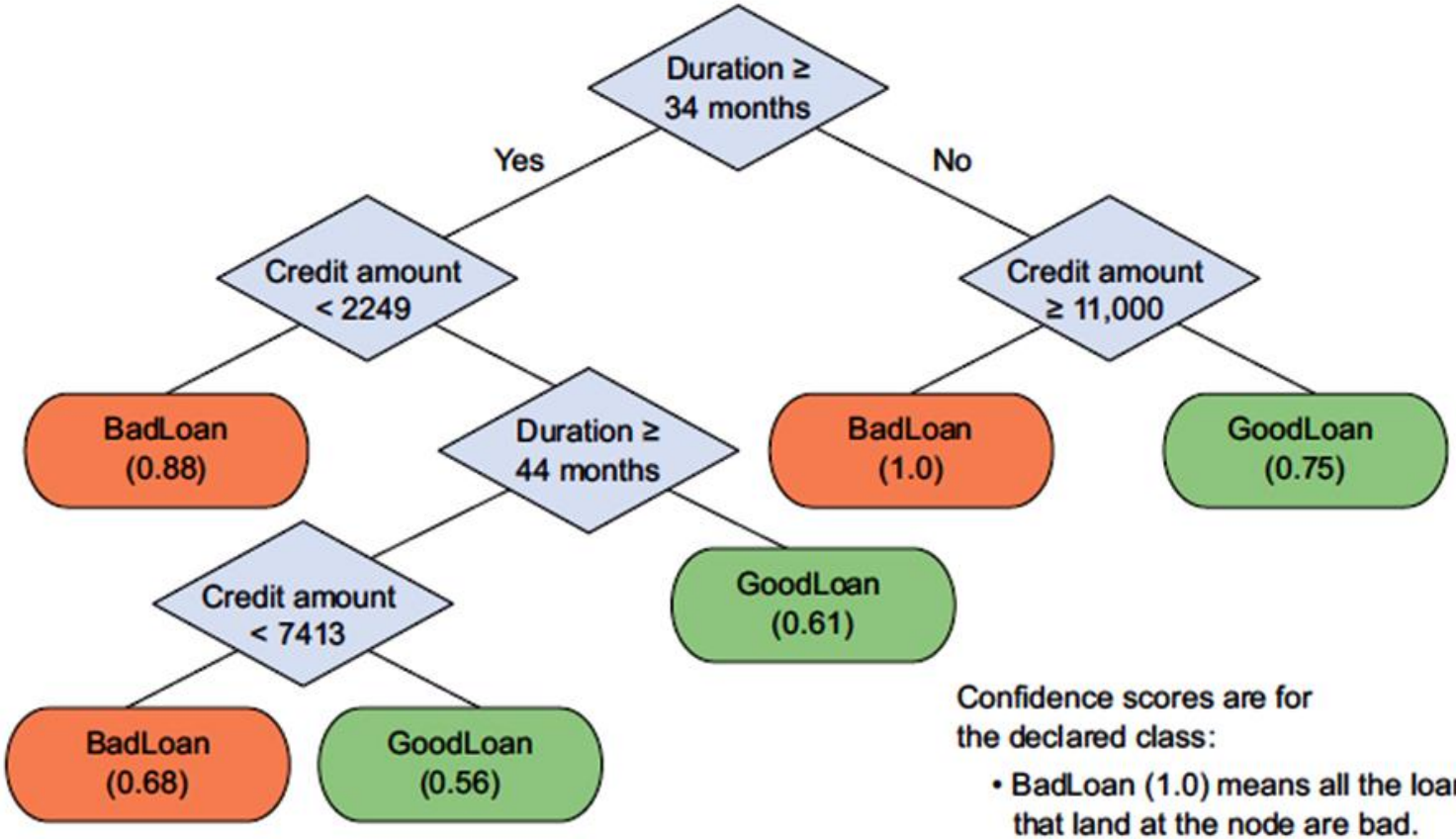
Some Common Classification Methods

Method	Description
Naïve Bayes	especially useful for problems with many categorical input variables, with a very large number of possible values, and text classification. Naive Bayes would be a good first attempt at solving the product categorization problem.
Decision Tree	useful when input variables interact with the output in “if-then” kinds of ways (such as IF age > 65, THEN insurance=T) or when input variables are redundant or correlated. an important extension of decision trees is random forests

Some Common Classification Methods

Method	Description
Logistic Regression	appropriate when you want to estimate class probabilities in addition to class assignments. An example use of a logistic regression–based classifier is estimating the probability of fraud in credit card purchases.
Support vector machines (SVMs)	are useful when there are very many input variables or when input variables interact with the outcome or with each other in complicated (nonlinear) ways. SVMs make fewer assumptions about variable distribution than do many other methods

A decision tree model for finding bad loan applications



Confidence scores are for the declared class:

- BadLoan (1.0) means all the loans that land at the node are bad.
- GoodLoan (0.75) means 75% of the loans that land at the node are good.

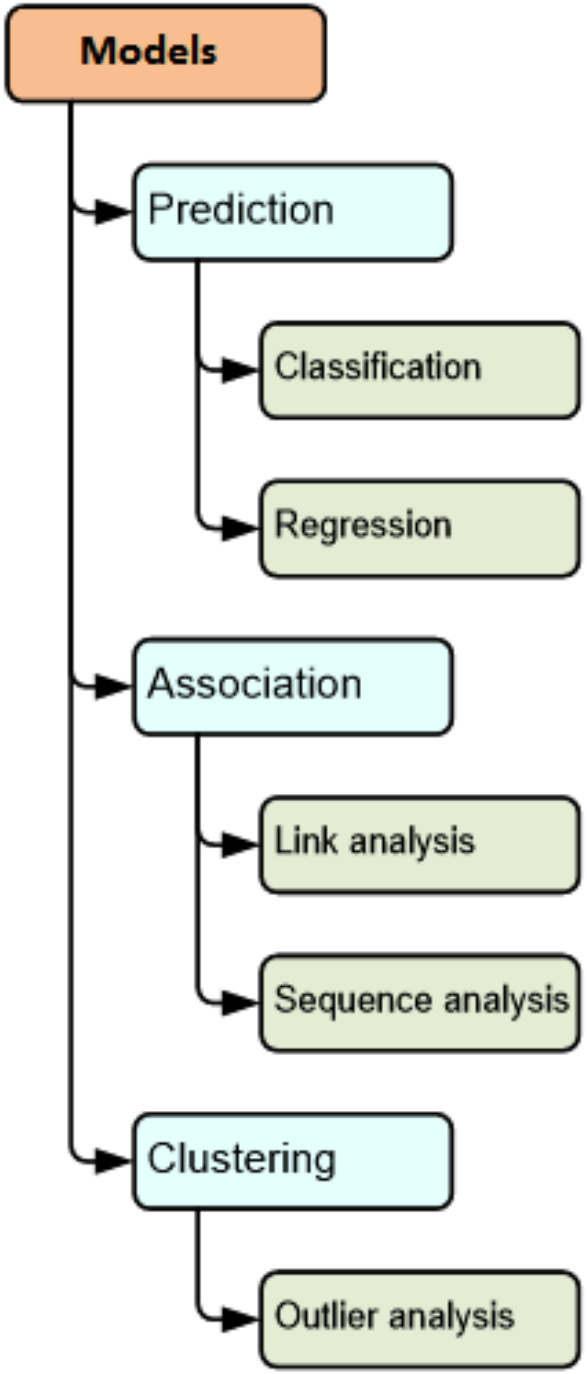
Splitting Data

Training Set
Testing Set
New Data

Types of Learning

Supervised
Un Supervised

Models in Data Science



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)