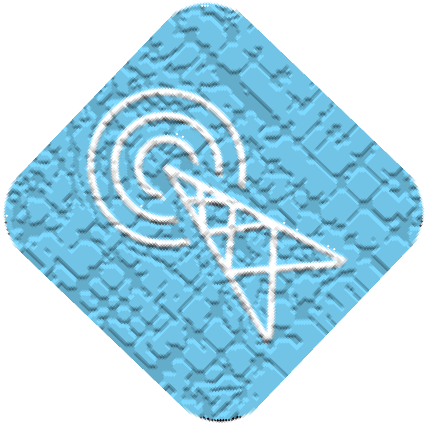
Introduction to Data Science

Overview

Objective

After completing this lesson you will be able to:



- Describe business analytics
- Explain the components of business analytics
- Explain the usage of business analytics in various domains

Business Analytics–Definition

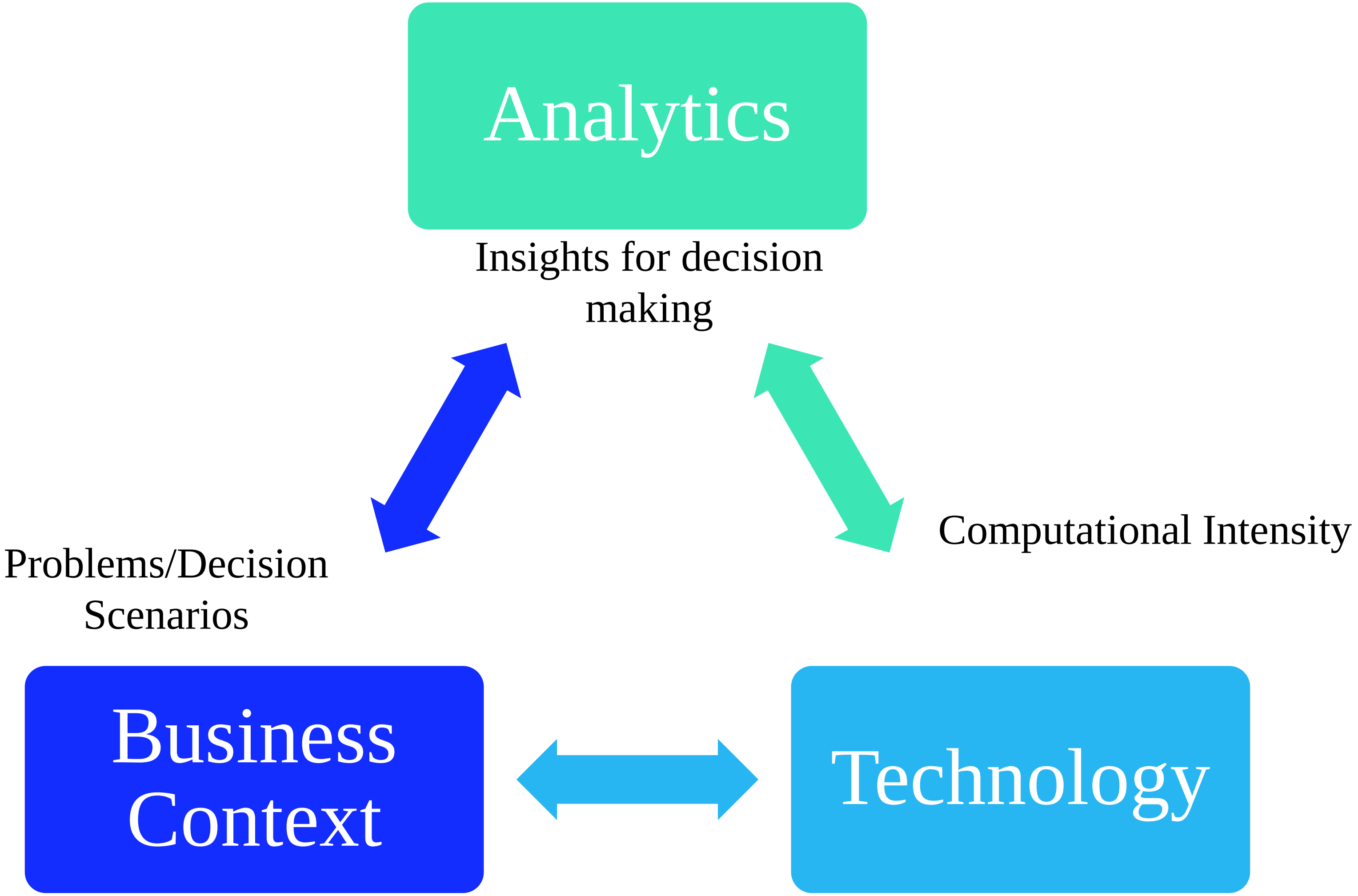
- Business analytics (BA) refers to the tools, techniques and processes for continuous exploration and investigation of past data to gain insights and help in decision making.
- Business Analytics is an integration between science, technology and business context that assist data driven decision making.

Data Explosion

- About 350 million photos are uploaded every day in the Facebook
- Amount of credit card debt in US: $762.1 billion
- Amount of credit card debt in India: Rs. 45,383 crore ($709 million)
- Loss due to global Credit card and debit card fraud $21.84 billion during 2015
- Every day, Walmart processes $36 million dollars an hour in sales
- BMTC with approx. 6000 buses plying in Bangalore sends 1 billion signals to the server updating its location every month

Interesting Stats: http://expandedramblings.com/

Analytics Trilogy



Analytics

Insights for decision making

Problems/Decision Scenarios

Computational Intensity

Business Context

Technology

Analytics in Use–Flipkart

- Forecast demand for each SKU.
- Predict customer cancellations and returns.
- Predict customer contacts at the customer service.
- Predict what a customer is likely to purchase in the future?
- How to optimize the delivery system?

The Game Changers

- Google
  - Used Markov chains to rank pages.
- Proctor and Gamble
  - Analytics as competitive strategy.
- Target
  - Predicts customer pregnancy.
- Capital One
  - Identifies the most profitable customer.
- Netflix:
  - Predicts movie ratings by customers (RMSE is 1%).
- Amazon.com:
  - 35% of sales come from product recommendations.

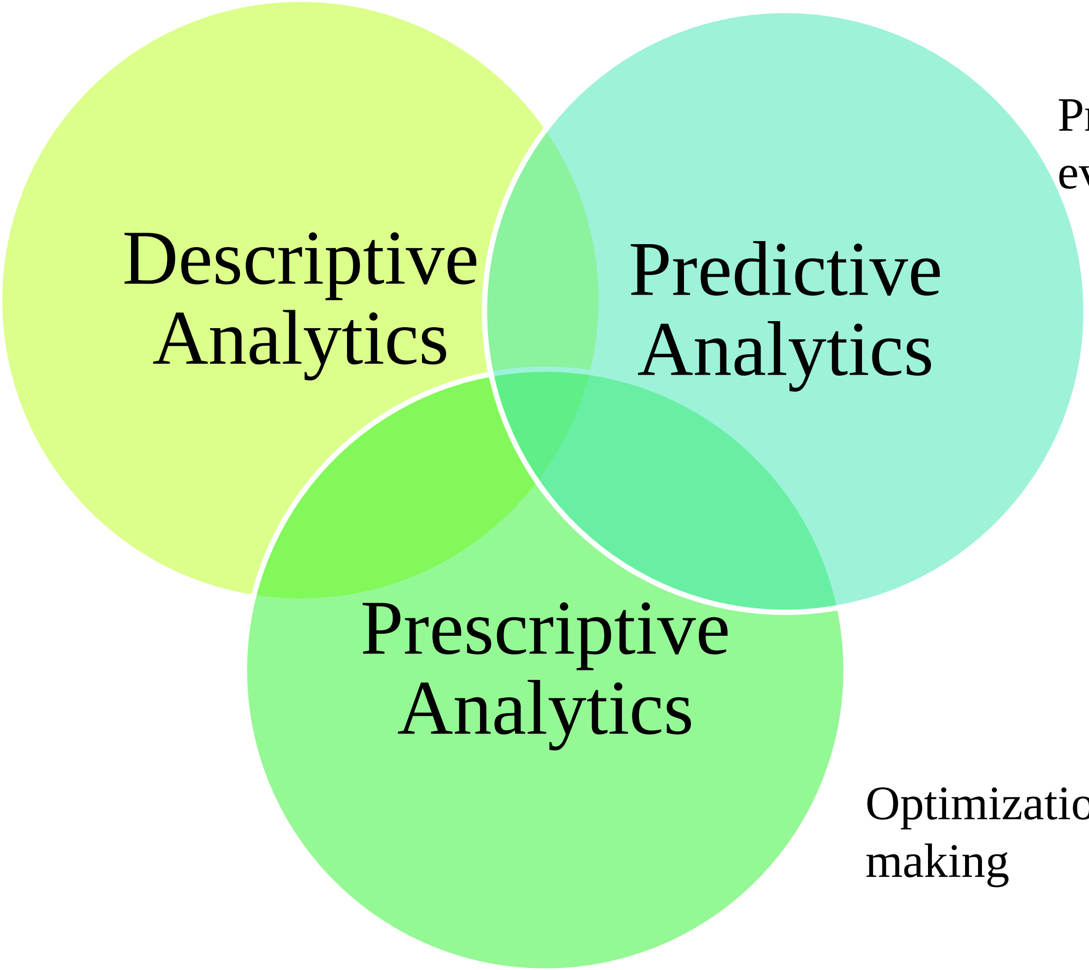
Data Scientists will be the sexiest job of 21st century

Harvard Business Review 2012

Components of Business Analytics

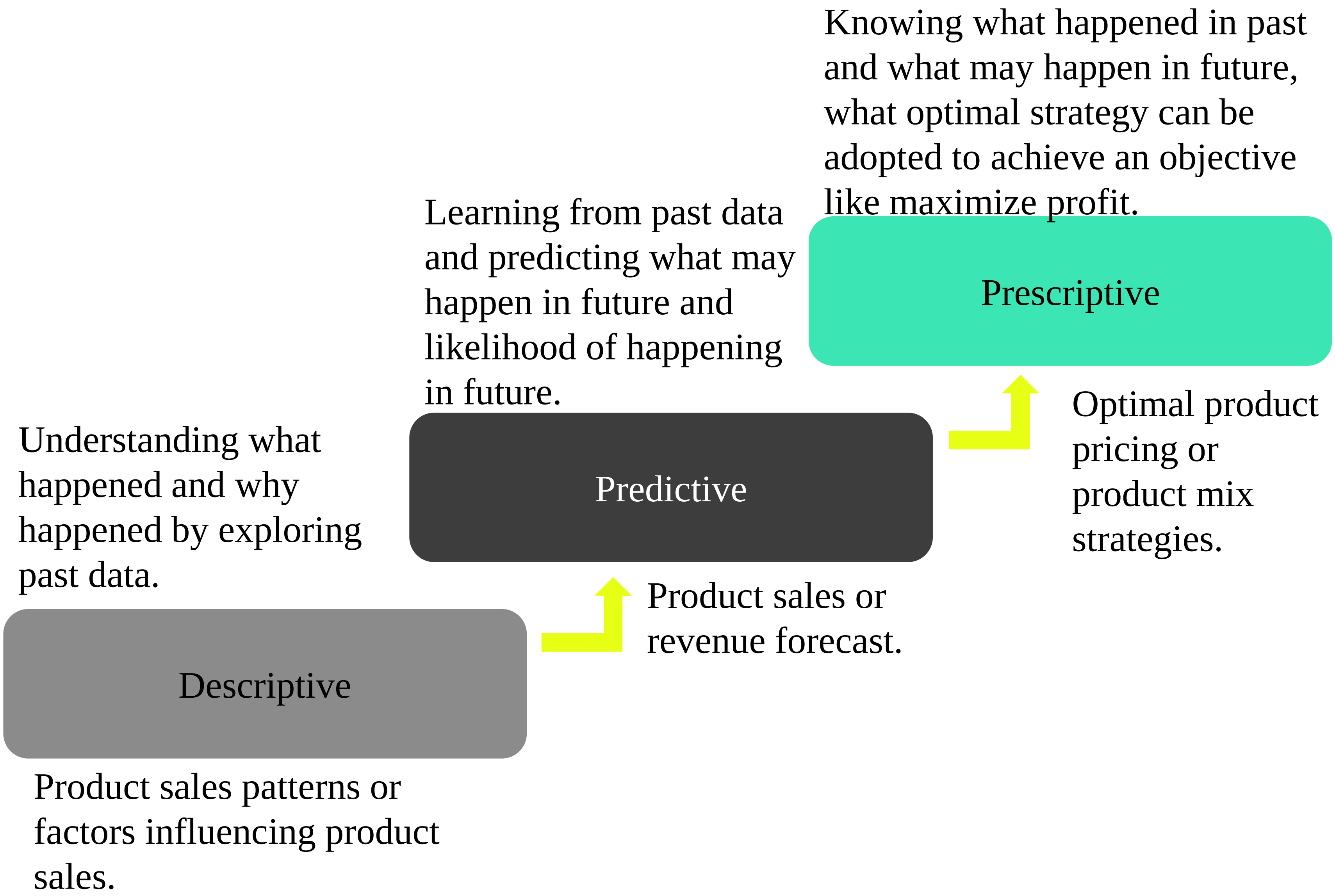
Data synthesis and Visualization

Predicting future events



Optimization and decision making

Components of Business Analytics



Understanding what happened and why happened by exploring past data.

Descriptive

Product sales patterns or factors influencing product sales.

Learning from past data and predicting what may happen in future and likelihood of happening in future.

Predictive

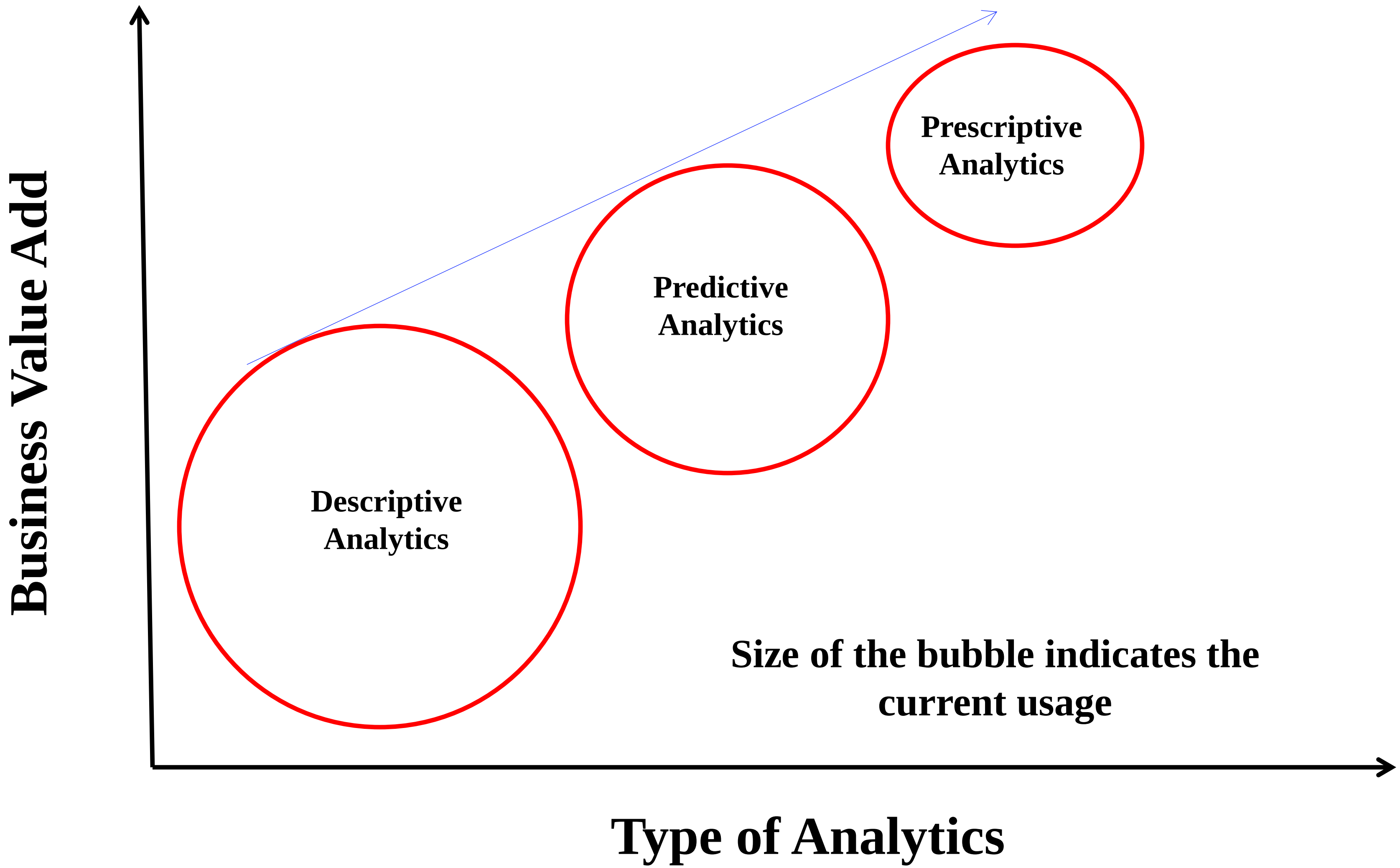
Product sales or revenue forecast.

Knowing what happened in past and what may happen in future, what optimal strategy can be adopted to achieve an objective like maximize profit.

Prescriptive

Optimal product pricing or product mix strategies.

Business Analytics & Intelligence



Size of the bubble indicates the current usage

Power of Descriptive Analytics

Descriptive Analytics Applications

- Most shoppers turn towards right when they enter the a retail store.
- Conversion rate of women shoppers is higher than male shoppers among electronic gadgets purchasers (Radio Shack).
- Strawberry pop-tarts sell 7 times more during hurricane compared to regular period (Wal Mart).
- Women car buyers prefer women sales person.

Broad Classification in Predictive Analytics

- Supervised Learning
  - Input (X’s) and Output (Y) both are known features

- Unsupervised Learning
  - Input (X’s) is known but Output (Y) is unknown

Predictive Analytics Application

- Which product the customer is likely to buy in his next purchase (recommender system).
- Which customer is likely to default in his/her loan payment.
- Who is likely to cancel the product that was ordered through e-commerce portal.

Prescriptive Analytics Application

- What is the optimal route for a delivery truck.
- Whether a company should introduce a new product?
- What is the optimal product mix?
- How to manage the fleet of vehicles owned by a company for employee drop and pick up?

Framework For Decision Making

Opportunity Identification

- Domain knowledge is very important at this stage of the analytics project. This will be a major challenge for many companies who do not know the capabilities of analytics.

Collection of relevant data

- Once the problem is defined clearly, the project team should identify and collect the relevant data. This may be an interactive process since "relevant data" may not be known in advance in many analytics projects. The existence of ERP systems will be very useful at this stage.

Data Pre-processing

- This would include data imputation and the creation of additional variables such as interaction variables and dummy variables in the case of predictive analytics projects.

Model Building

- Analytics model building is an iterative process that aims to find the best model. Several analytical tools and solution procedures will be used to find the best analytical model in this stage.

Communication of the data analysis

- The communication of the analytics output to the top management and clients plays a crucial role. Innovative data visualization techniques may be used in this stage.

Industry Wide Application of Analytics

Manufacturing

- *Supply chain analytics*
- *Quality and Process improvement*
- *Revenue and Cost Management*

Retail

- *Assortment Planning*
- *Promotion Planning*
- *Demand Forecasting*
- *Market Basket Analysis*
- *Customer Segmentation*

Healthcare

- *Clinical Care*
- *Hospitality related data*

Service

- *Demand Forecasting*
- *Service Quality Analysis*
- *Customer Segmentation*
- *Promotion*

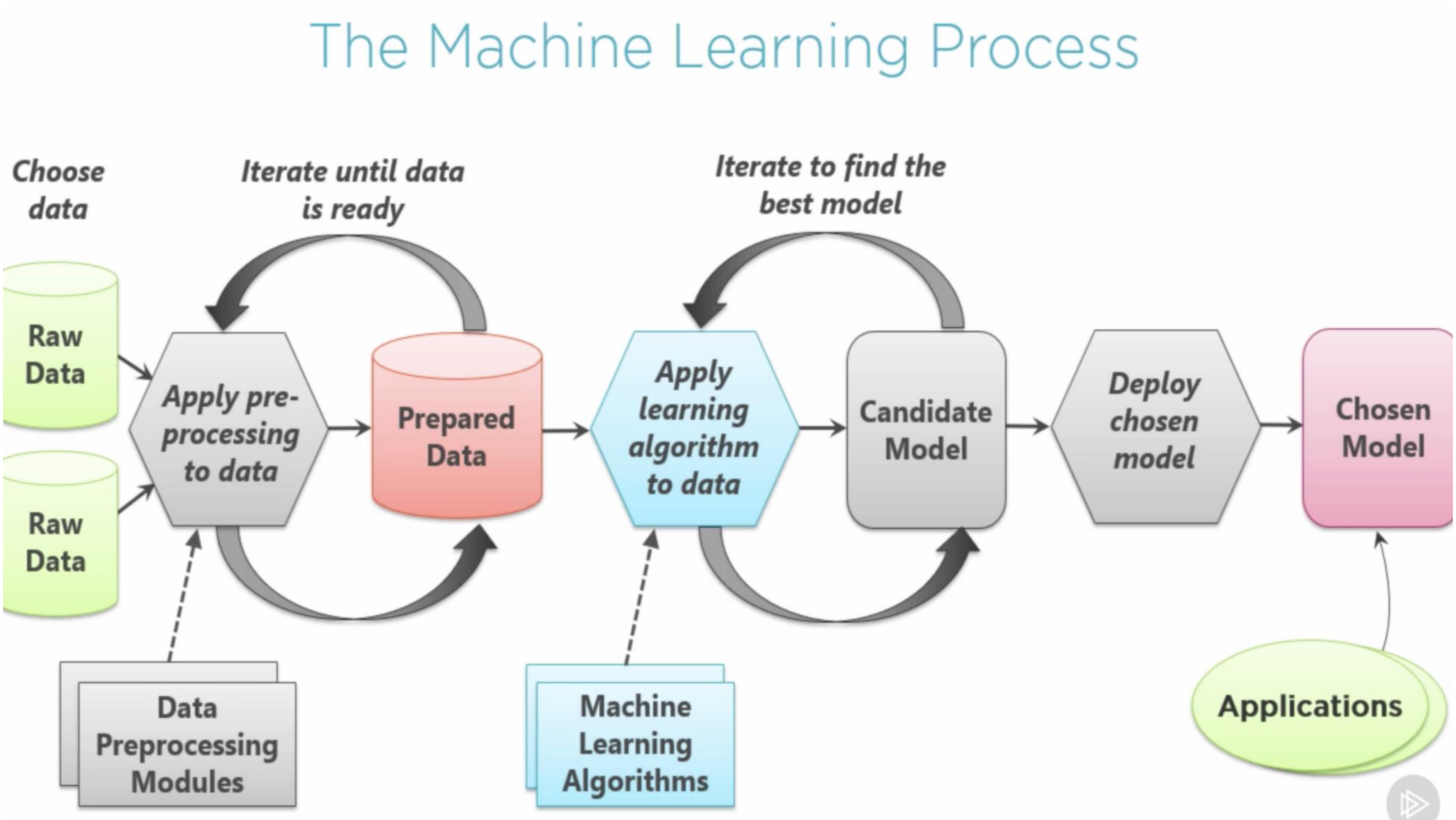
Banking and Finance

- *Service Demand Analysis*
- *Customer Transaction Analysis*
- *Credit Scoring*

IT and ITES (IT enabled services)

- *Demand for Analytics Services*
- *Software Development Cycle Time*

How to go about Machine Learning Path



The Machine Learning Process

Choose data

Iterate until data is ready

Iterate to find the best model

Raw Data

Raw Data

Apply pre-processing to data

Prepared Data

Apply learning algorithm to data

Candidate Model

Deploy chosen model

Chosen Model

Data Preprocessing Modules

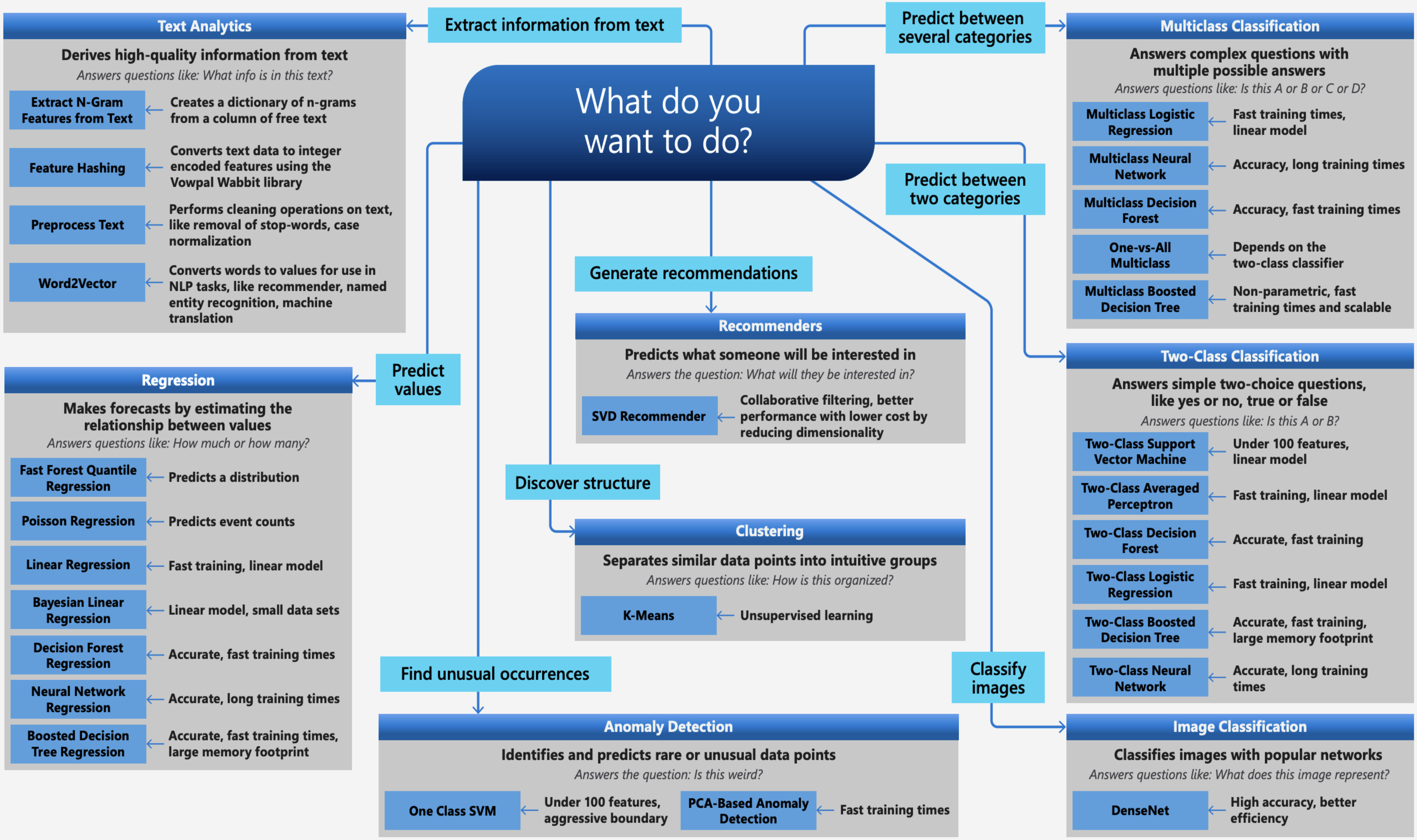
Machine Learning Algorithms

Applications

Source : https://volcanohong.github.io/2016/09/01/machine-learning-notes/

Microsoft Azure Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.



What do you want to do?

Extract information from text → Text Analytics

Predict values → Regression

Generate recommendations → Recommenders

Discover structure → Clustering

Find unusual occurrences → Anomaly Detection

Predict between several categories → Multiclass Classification

Predict between two categories → Two-Class Classification

Classify images → Image Classification

Text Analytics

Derives high-quality information from text

Answers questions like: What info is in this text?

Extract N-Gram Features from Text	Creates a dictionary of n-grams from a column of free text
Feature Hashing	Converts text data to integer encoded features using the Vowpal Wabbit library
Preprocess Text	Performs cleaning operations on text, like removal of stop-words, case normalization
Word2Vector	Converts words to values for use in NLP tasks, like recommender, named entity recognition, machine translation

Regression

Makes forecasts by estimating the relationship between values

Answers questions like: How much or how many?

Fast Forest Quantile Regression	Predicts a distribution
Poisson Regression	Predicts event counts
Linear Regression	Fast training, linear model
Bayesian Linear Regression	Linear model, small data sets
Decision Forest Regression	Accurate, fast training times
Neural Network Regression	Accurate, long training times
Boosted Decision Tree Regression	Accurate, fast training times, large memory footprint

Recommenders

Predicts what someone will be interested in

Answers the question: What will they be interested in?

SVD Recommender	Collaborative filtering, better performance with lower cost by reducing dimensionality

Clustering

Separates similar data points into intuitive groups

Answers questions like: How is this organized?

K-Means	Unsupervised learning

Anomaly Detection

Identifies and predicts rare or unusual data points

Answers the question: Is this weird?

One Class SVM	Under 100 features, aggressive boundary
PCA-Based Anomaly Detection	Fast training times

Multiclass Classification

Answers complex questions with multiple possible answers

Answers questions like: Is this A or B or C or D?

Multiclass Logistic Regression	Fast training times, linear model
Multiclass Neural Network	Accuracy, long training times
Multiclass Decision Forest	Accuracy, fast training times
One-vs-All Multiclass	Depends on the two-class classifier
Multiclass Boosted Decision Tree	Non-parametric, fast training times and scalable

Two-Class Classification

Answers simple two-choice questions, like yes or no, true or false

Answers questions like: Is this A or B?

Two-Class Support Vector Machine	Under 100 features, linear model
Two-Class Averaged Perceptron	Fast training, linear model
Two-Class Decision Forest	Accurate, fast training
Two-Class Logistic Regression	Fast training, linear model
Two-Class Boosted Decision Tree	Accurate, fast training, large memory footprint
Two-Class Neural Network	Accurate, long training times

Image Classification

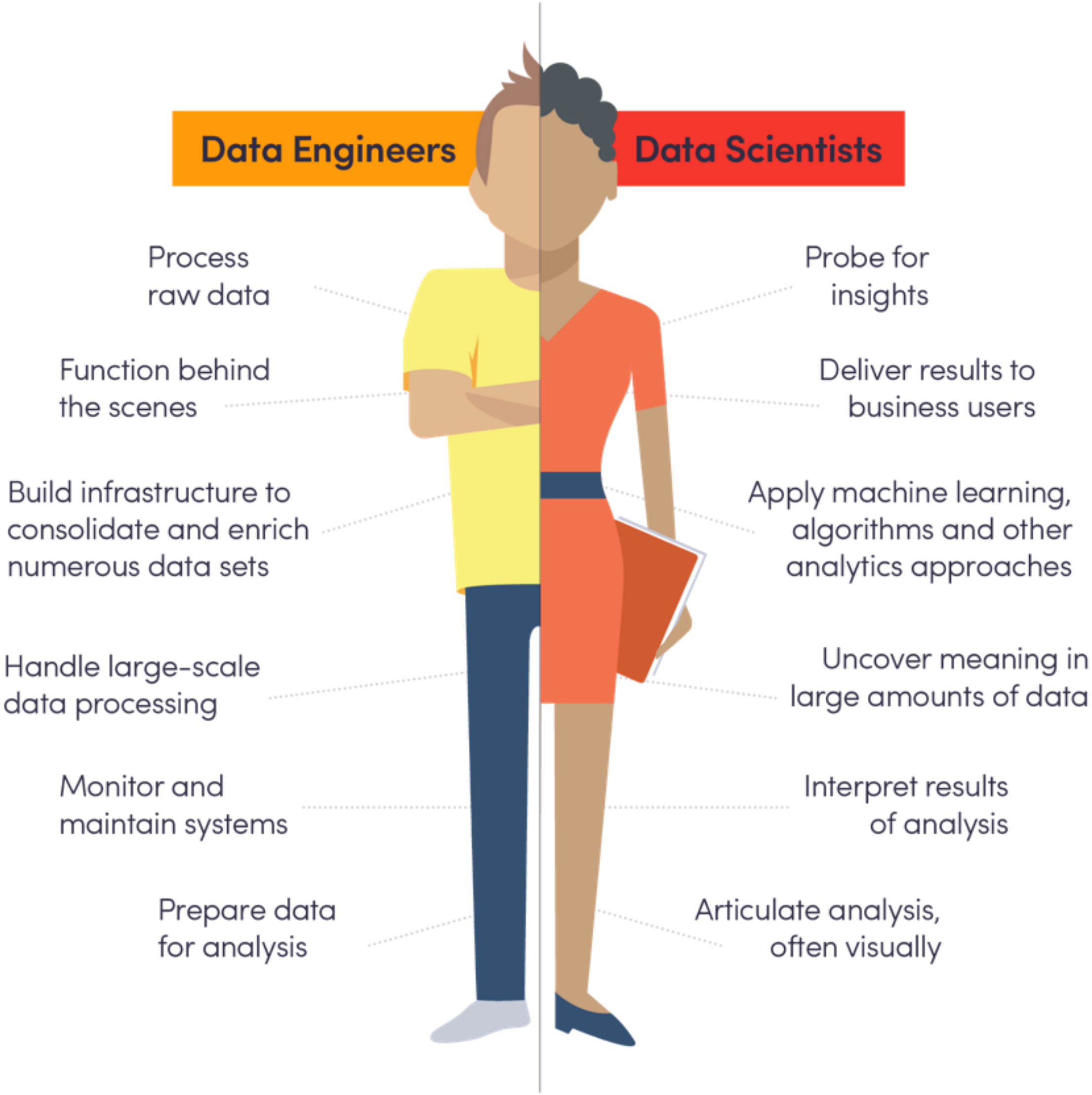
Classifies images with popular networks

Answers questions like: What does this image represent?

DenseNet	High accuracy, better efficiency

Share this poster: aka.ms/mlcheatsheet

Microsoft


Data Engineers
Data Scientists
Process raw data
Function behind the scenes
Build infrastructure to consolidate and enrich numerous data sets
Handle large-scale data processing
Monitor and maintain systems
Prepare data for analysis
Probe for insights
Deliver results to business users
Apply machine learning, algorithms and other analytics approaches
Uncover meaning in large amounts of data
Interpret results of analysis
Articulate analysis, often visually

Source: https://towardsdatascience.com/data-engineering-the-cousin-of-data-science-is-troublesome-3a9332b532ae

What Tools are available?

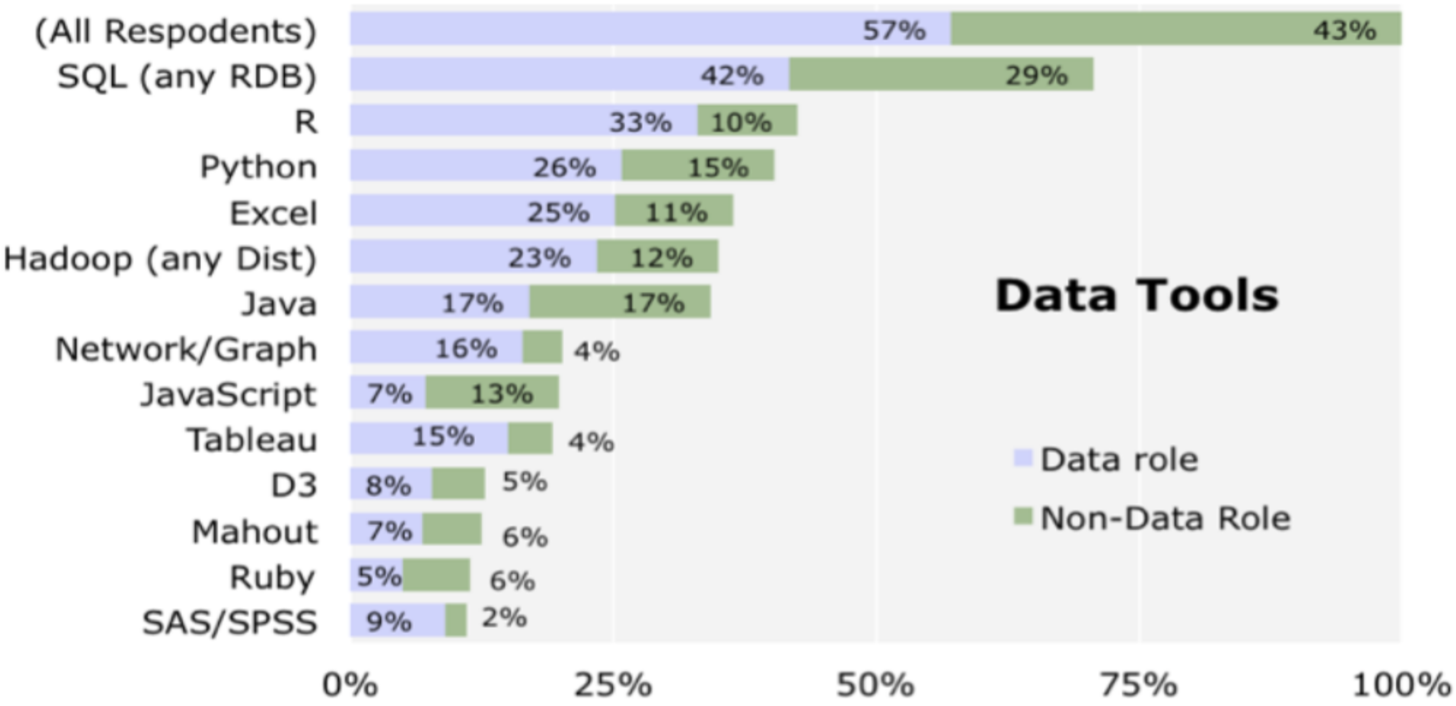
R Vs. Python

R	Python
Built for Statistical Analysis.	General Purpose Language. Main objective is productivity and readability.
Primarily used in academics and research. Enterprise have started adopting it for analysis.	Has a very strong presence in enterprises for large number of software developments. Easier adoption in enterprises as strong development experience already exists.
Integration with other enterprise systems are not straightforward.	Integration with other enterprise systems or applications are easier.

Python

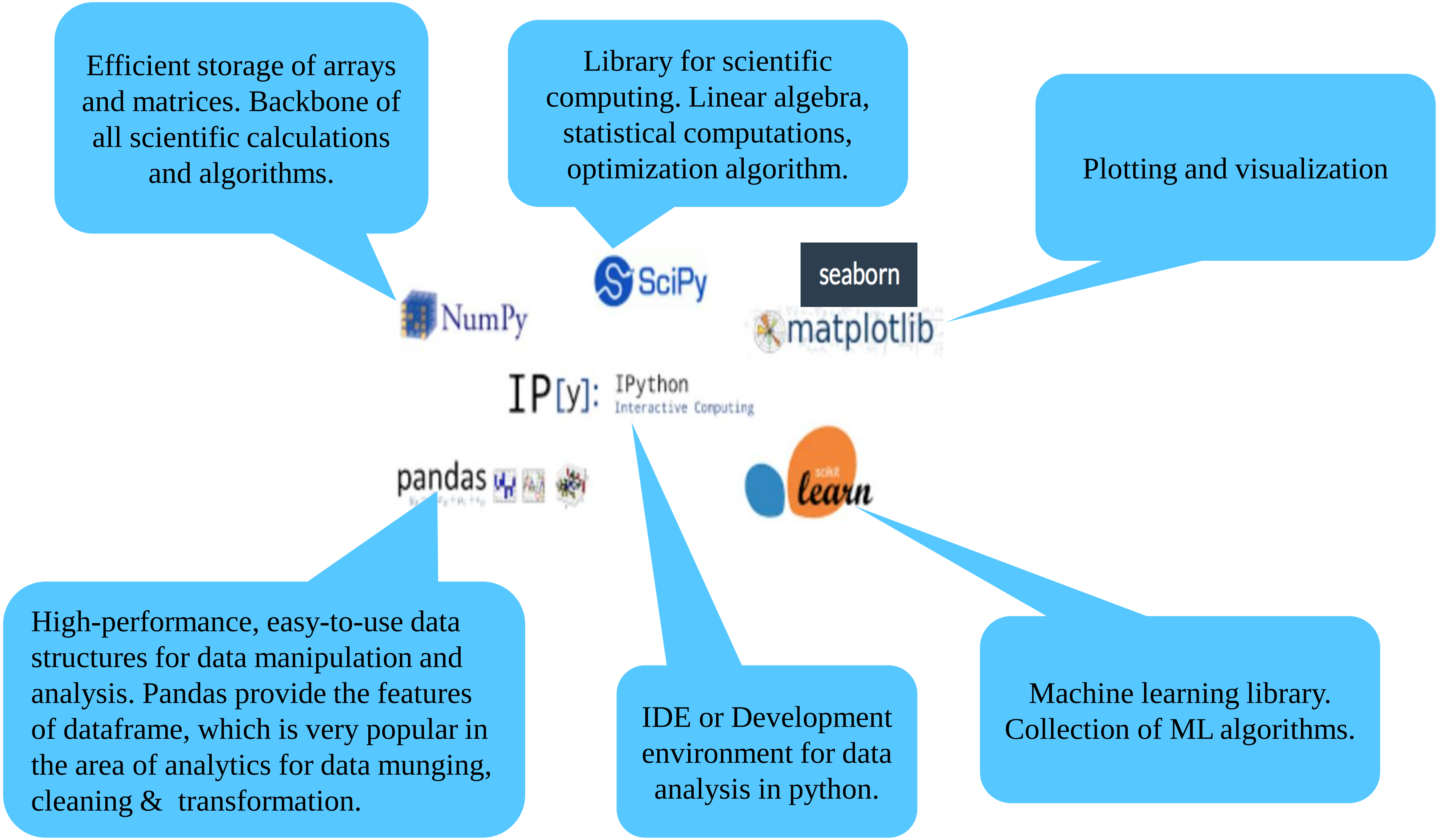
- Multi-purpose
  - Web Developments
  - Scripting
  - Server Side Developments
  - Statistical Learnings & Machine Learnings
- Object Oriented
- Interpreted
- Strongly typed and Dynamically typed
- Focus on readability and productivity

Python Stack For Data Science



http://blog.revolutionanalytics.com/2014/01/in-data-scientist-survey-r-is-the-most-used-tool-other-than-databases.html

Python Stack For Data Science



Efficient storage of arrays and matrices. Backbone of all scientific calculations and algorithms.

Library for scientific computing. Linear algebra, statistical computations, optimization algorithm.

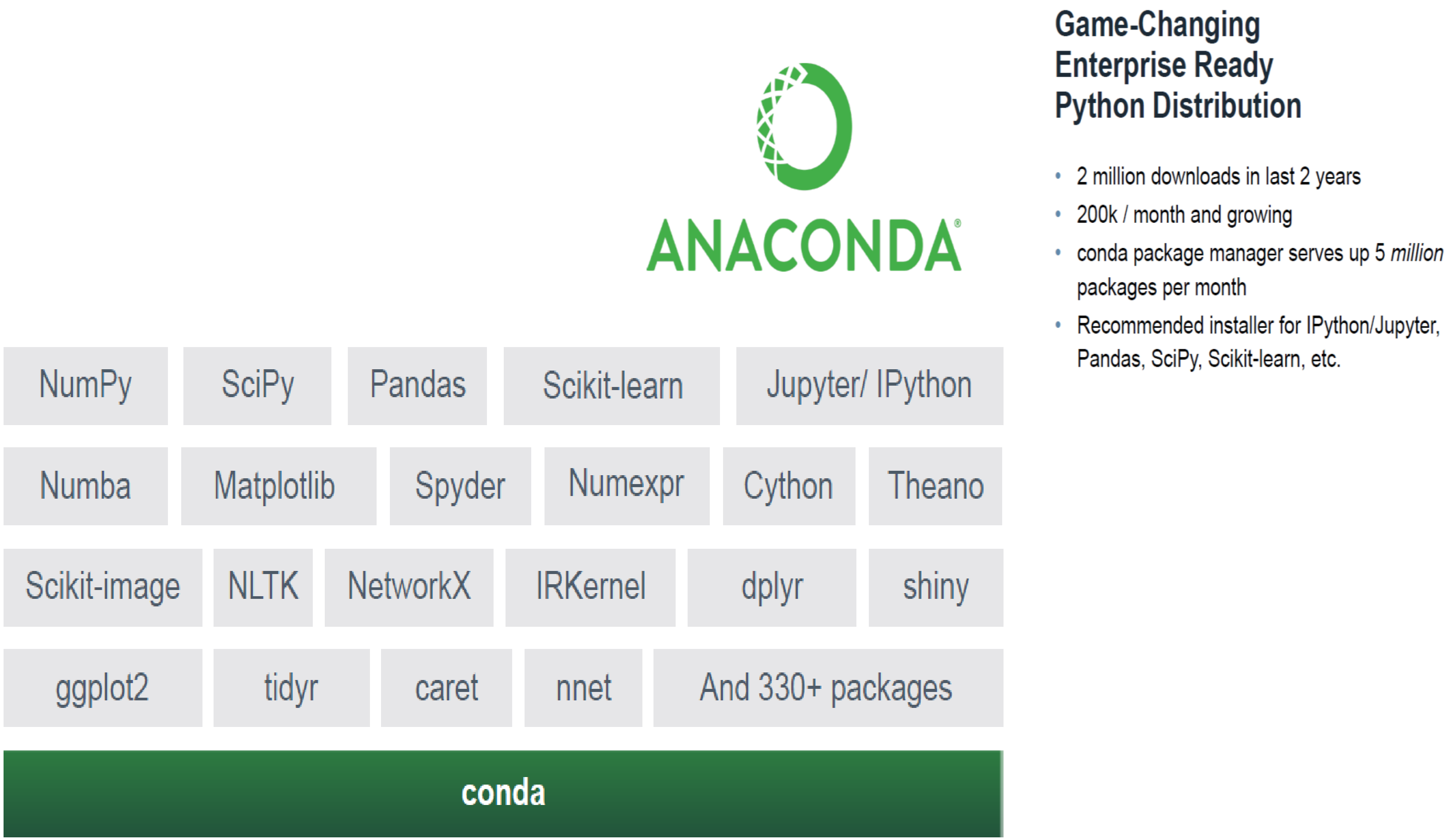
Plotting and visualization

High-performance, easy-to-use data structures for data manipulation and analysis. Pandas provide the features of dataframe, which is very popular in the area of analytics for data munging, cleaning & transformation.

IDE or Development environment for data analysis in python.

Machine learning library. Collection of ML algorithms.

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.

Download link: https://www.continuum.io/downloads

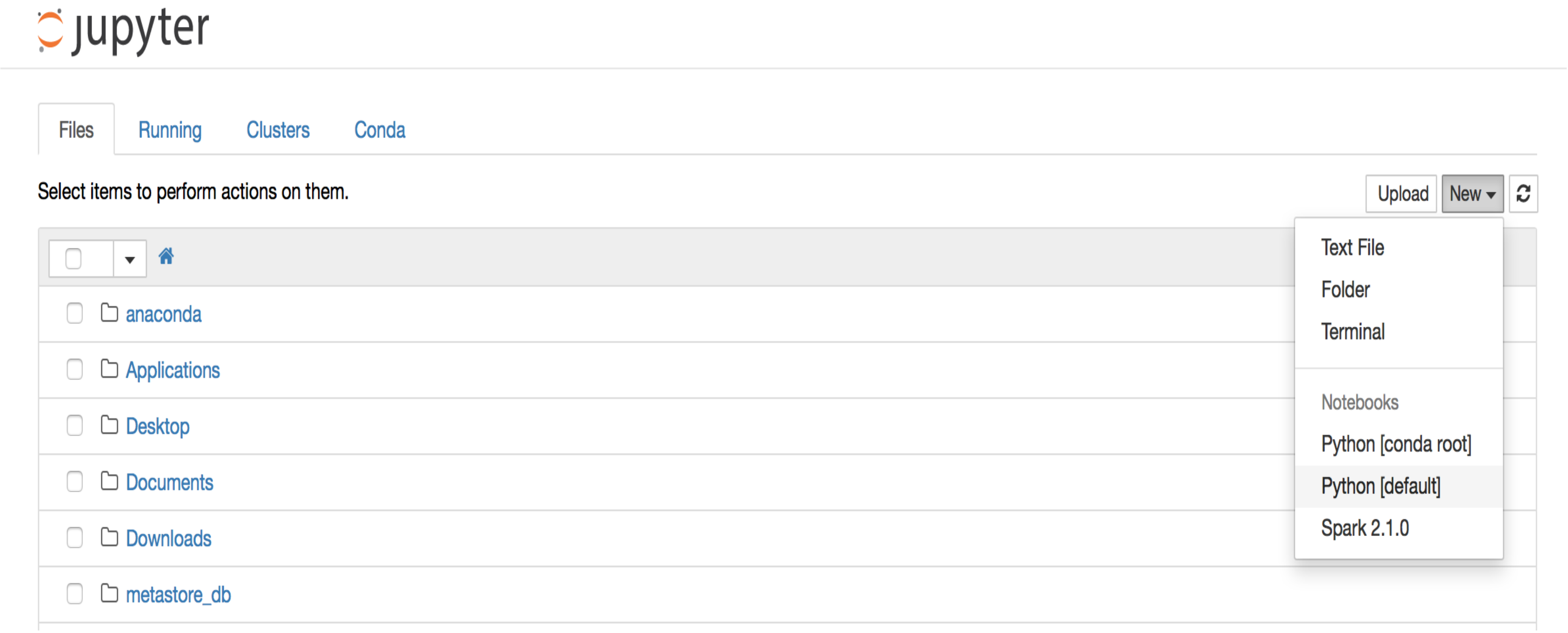
Source: Continuum Analytics

Start Jupyter notebook

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

- For Windows
  - Go to Anaconda command prompt and enter
    - **jupyter notebook --ip=***

Start a jupyter notebook



jupyter

Files Running Clusters Conda

Select items to perform actions on them.

Upload New ▾

- anaconda
- Applications
- Desktop
- Documents
- Downloads
- metastore_db

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

Numpy and Pandas

NumPy

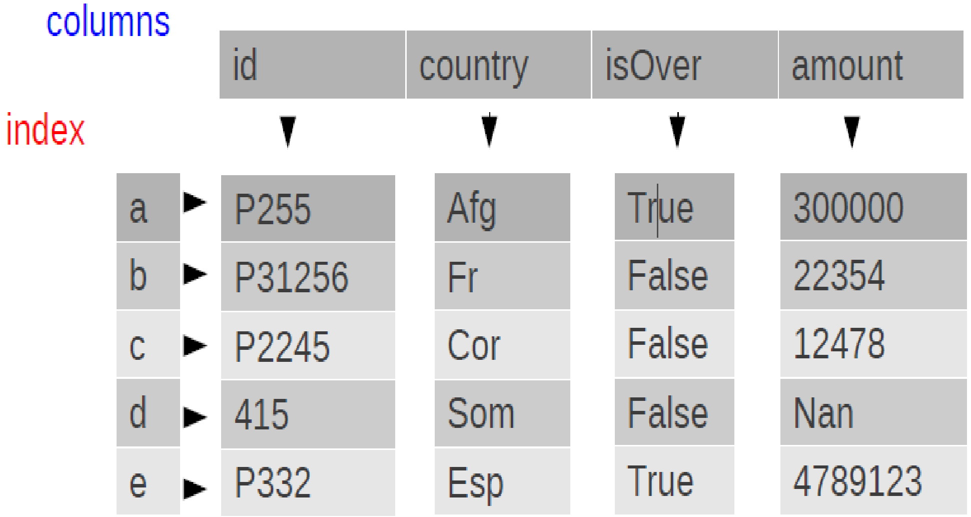
- Library for mathematical and numerical routines like Matlab
- Provides basic routines
  - Manipulating large arrays and matrices of numeric data.
- Foundational library for all statistical and machine learnings
  - Pandas and SciPy
- Using NumPy library

  import numpy as np

Pandas

- Recent API based on Numpy, Optimized for performance
- Easy to work with messy and irregularly indexed data
- Adopts concepts of R language dataframes
- The two basics structures of pandas
  - Series 1d array
  - DataFrame 2d array
- Typical Data Munging Activities
  - Filtering, selecting data
  - Aggregating, transforming data
  - Joining, concatenating, merging data
  - Descriptive basics statistics

Pandas


columns
index

	id	country	isOver	amount
a	P255	Afg	True	300000
b	P31256	Fr	False	22354
c	P2245	Cor	False	12478
d	415	Som	False	Nan
e	P332	Esp	True	4789123

Table like structure

- 2D data structure
- Row and column index
- Size mutable: insert or delete columns
- SQL like transformations – select, groupby, aggregations, filtering, joining etc.

Summary

Summary of the topics covered in this lesson:

- With the data explosion across industry, the usage of analytics in decision making will become the most critical factor for being competitive in business.
- Descriptive analytics becomes the stepping stone to all the complex problems which can be solved using analytics.

End of Lesson–Introduction to Business Analytics