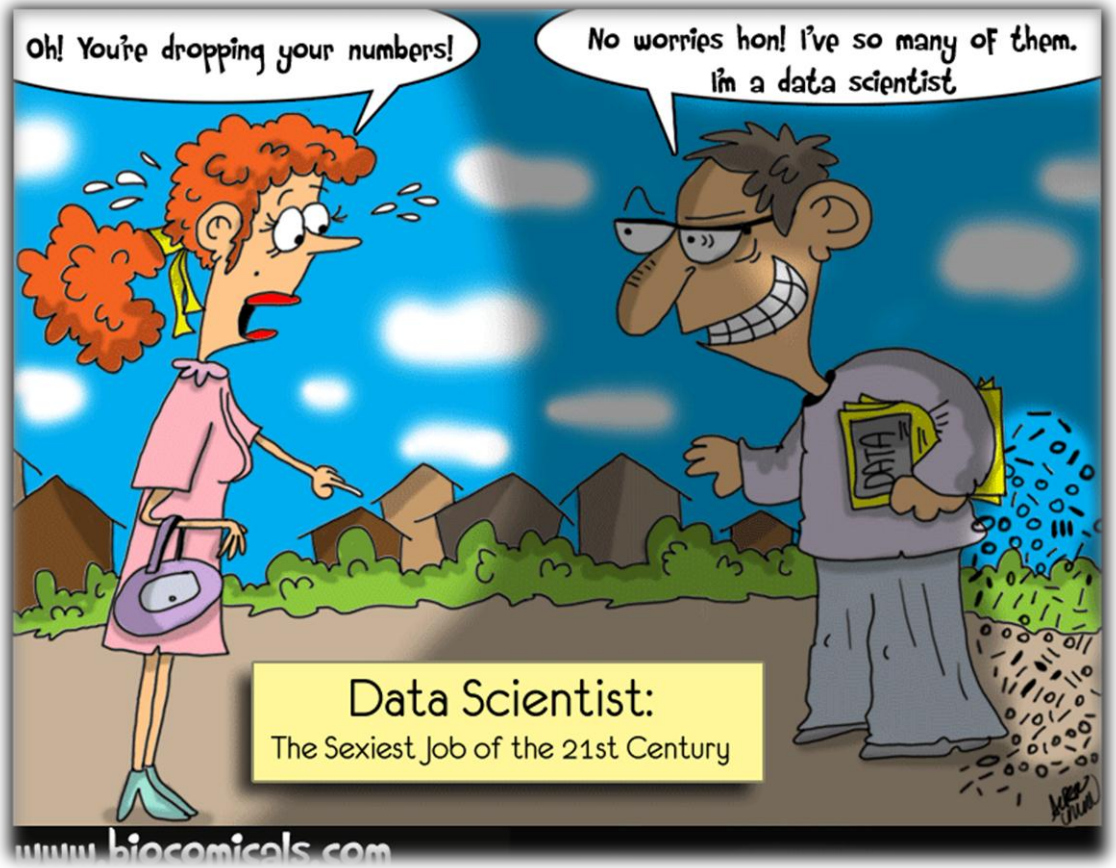
DataInquest

Big Data Technologies and Data Science training for all!!

Data Science

Data Science



Big Data

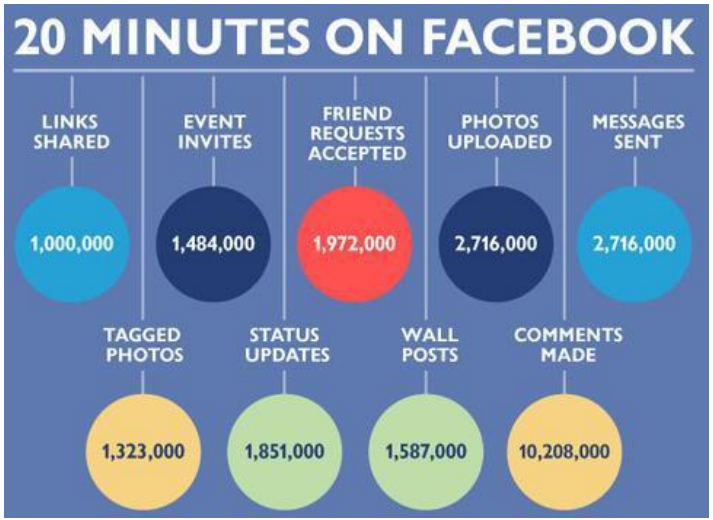
What is Big Data?

- ✓ Lots of Data (Terabytes or Petabytes)
- ✓ Systems / Enterprises generate huge amount of data from Terabytes to and even Petabytes of information.

A airline jet collects 10 terabytes of sensor data for every 30 minutes of flying time.

NYSE generates about one terabyte of new trade data per day to Perform stock trading analytics to determine trends for optimal trades.

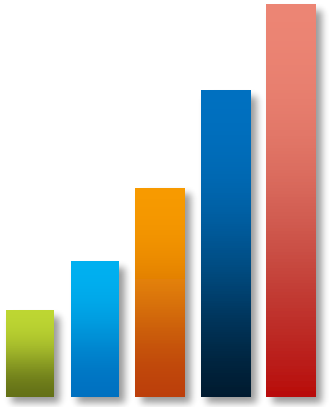
Facebook Example



- ✓ Facebook users spend **10.5 billion** minutes (almost 20,000 years) online on the social network.
- ✓ Facebook has an average of **3.2 billion** likes and comments are posted every day.

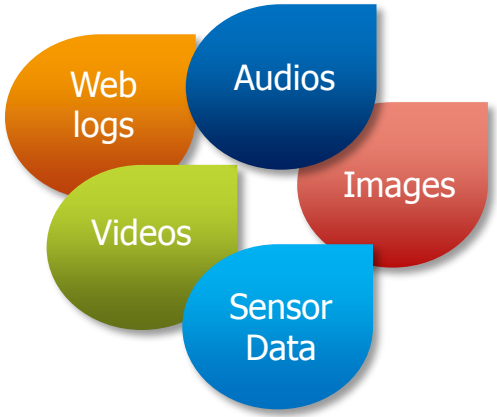
IBM's Definition

- ✓ **IBM's Definition – Big Data Characteristics**
  http://www-01.ibm.com/software/data/bigdata/



VOLUME

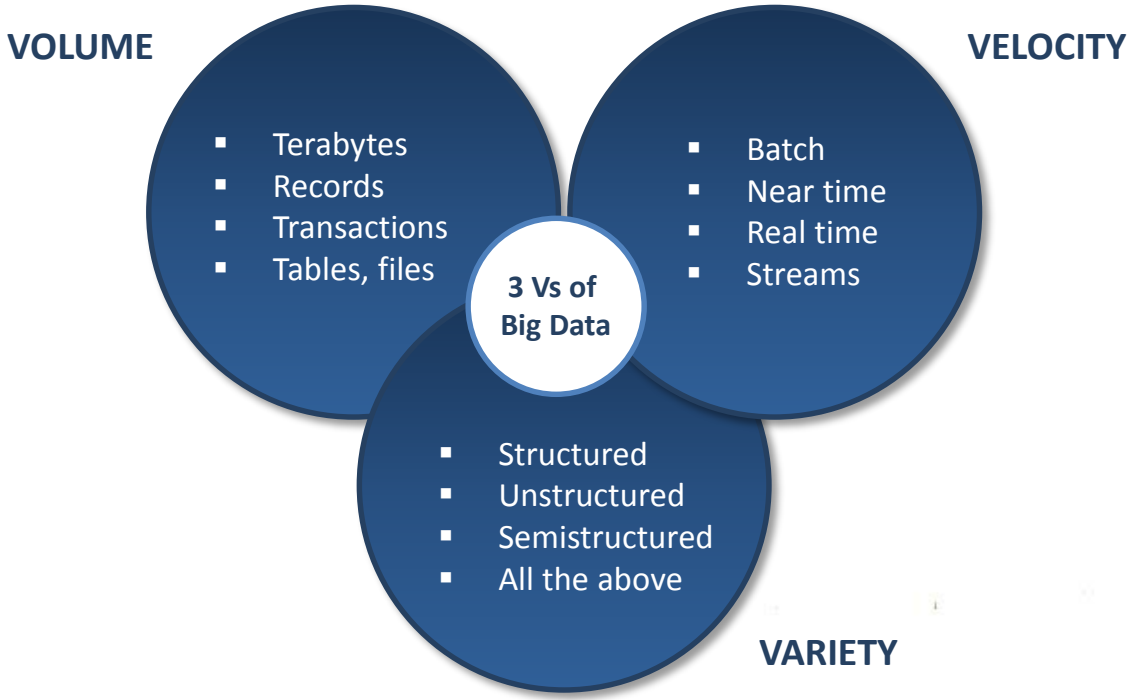
VELOCITY



VARIETY

IBM's Definition

- ✓ **IBM's Definition – Big Data Characteristics**
  http://www-01.ibm.com/software/data/bigdata/



Common Big Data Customer Scenarios (Contd.)

- ✓ Government
  - ✓ Fraud Detection And Cyber Security
  - ✓ Welfare schemes
  - ✓ Justice
- ✓ Healthcare & Life Sciences
  - ✓ Health information exchange
  - ✓ Gene sequencing
  - ✓ Serialization
  - ✓ Healthcare service quality improvements
  - ✓ Drug Safety

http://wiki.apache.org/hadoop/PoweredBy

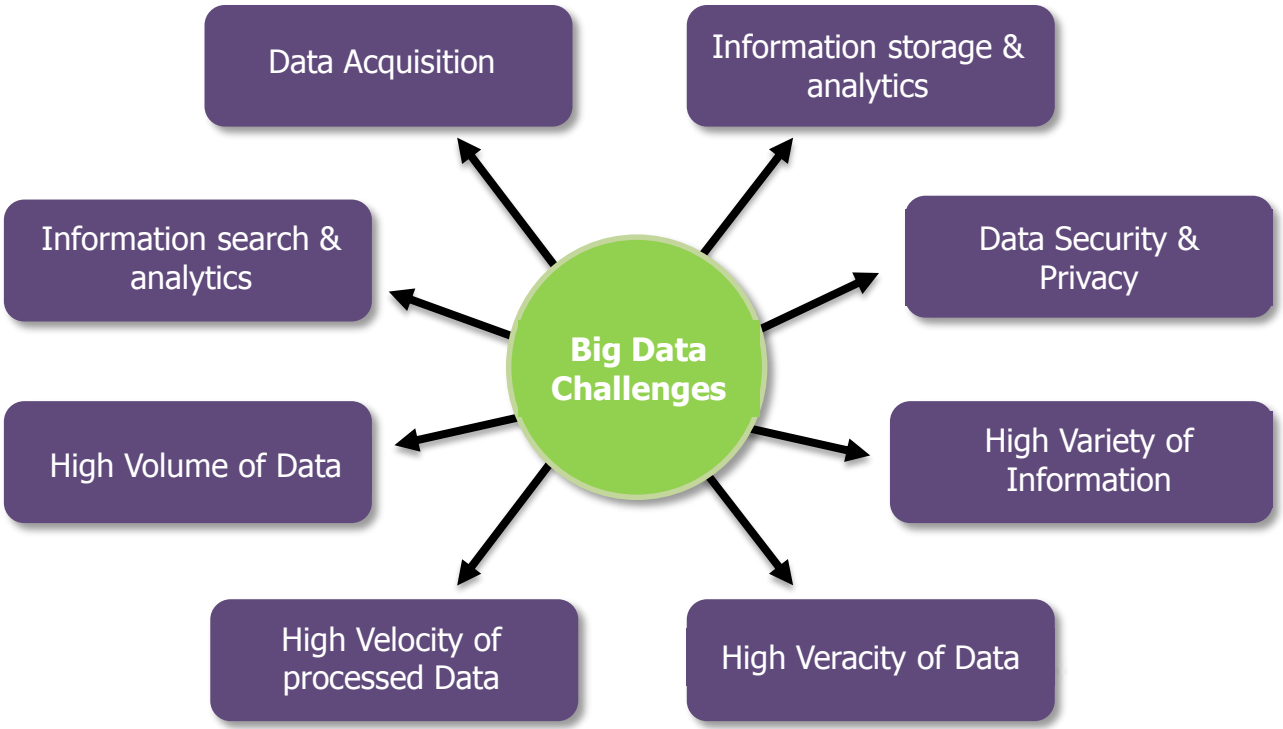
Common Big Data Customer Scenarios (Contd.)

- ✓ Banks and Financial services
  - ✓ Modeling True Risk
  - ✓ Threat Analysis
  - ✓ Fraud Detection
  - ✓ Trade Surveillance
  - ✓ Credit Scoring And Analysis

- ✓ Retail
  - ✓ Point of sales Transaction Analysis
  - ✓ Customer Churn Analysis
  - ✓ Sentiment Analysis

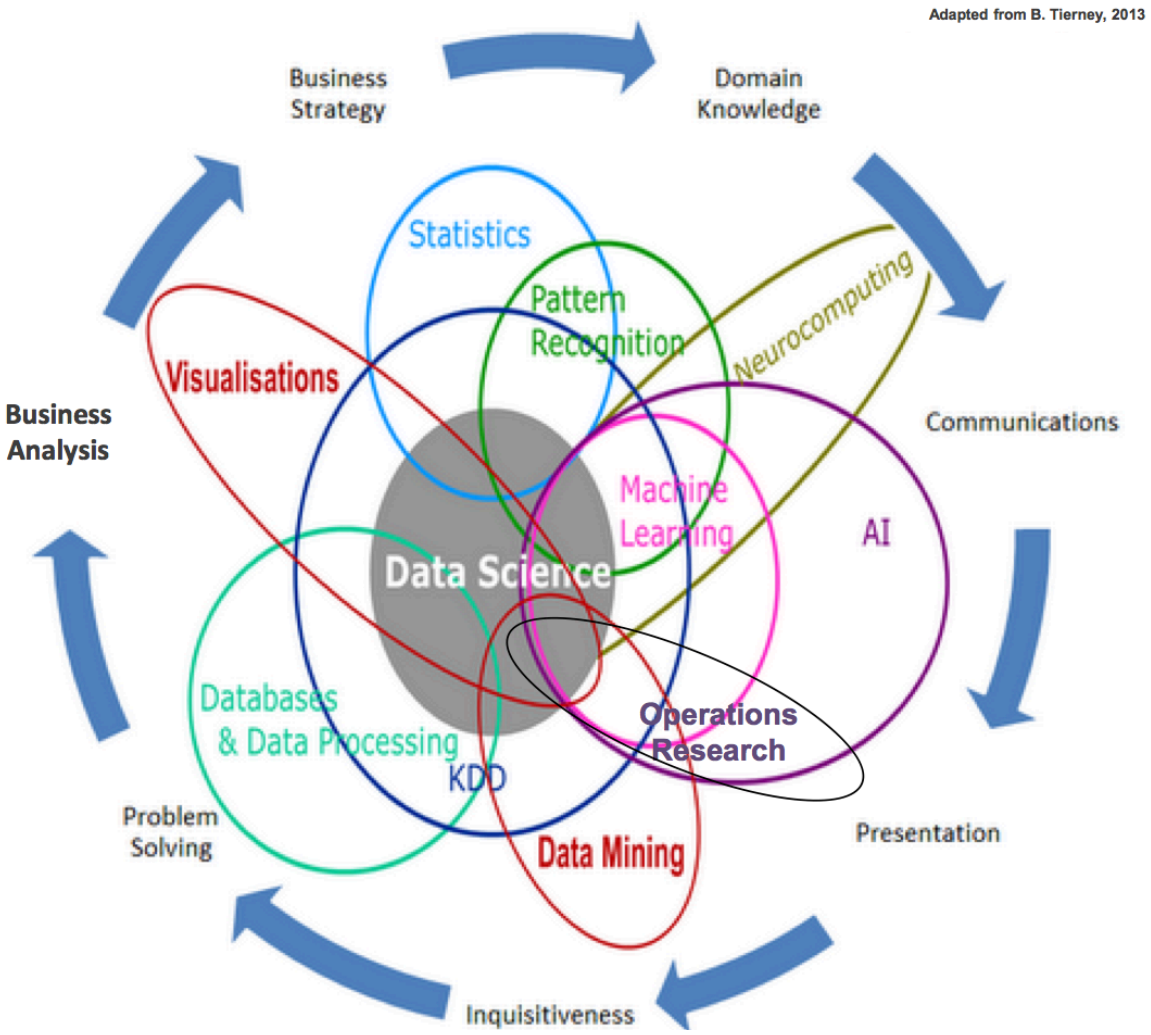
http://wiki.apache.org/hadoop/PoweredBy

Big Data: The Challenges



What is Data Science ?

Data Science in Multidisciplinary

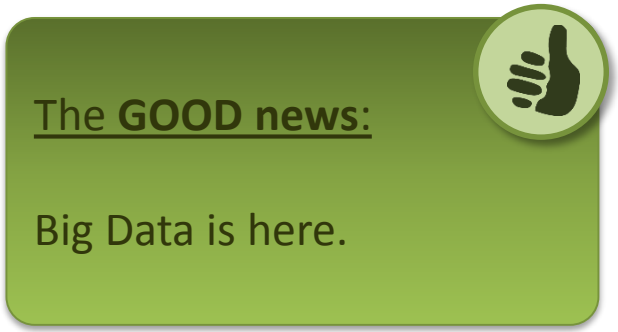


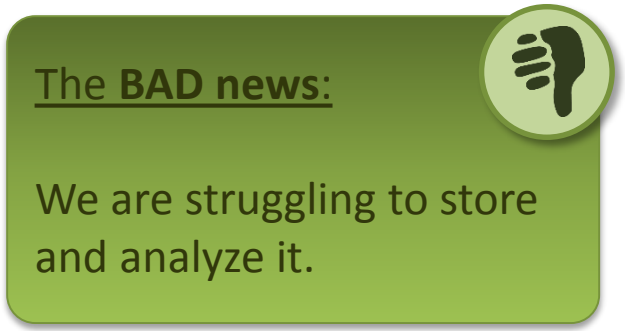
Data Science

“More data usually beats better algorithms,”

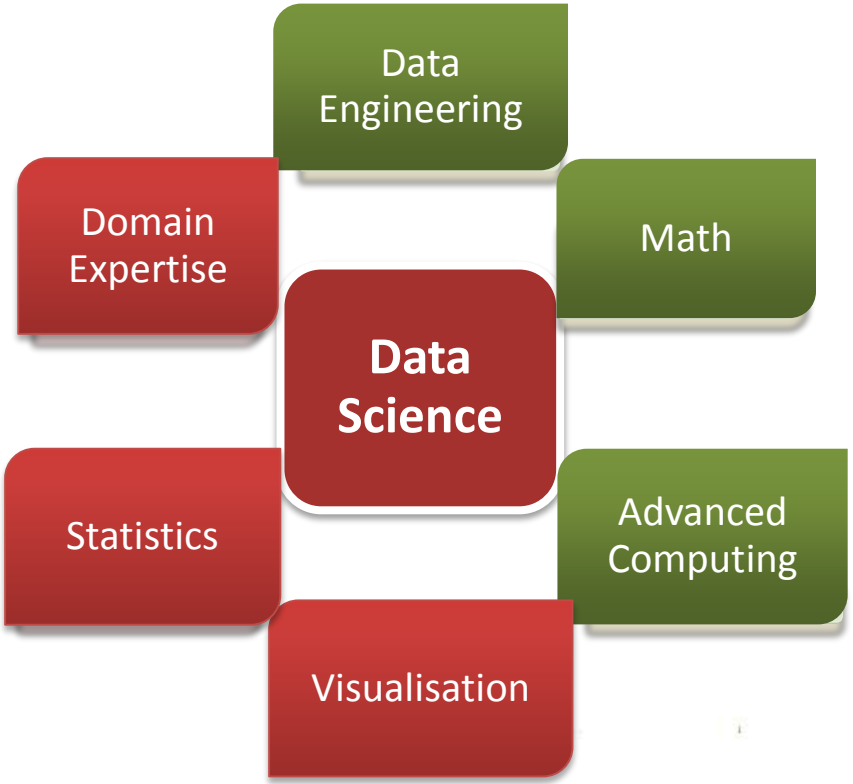
Such as: Recommending movies or music based on past preferences.

No matter how extremely unpleasant your algorithm is, , they can often be beaten simply by having more data (and a less sophisticated algorithm).





Data Science: Components

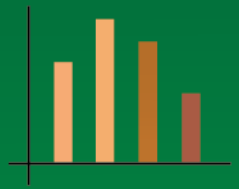


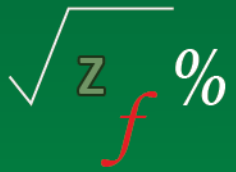
Data Science

WHAT IS DATA SCIENCE ?

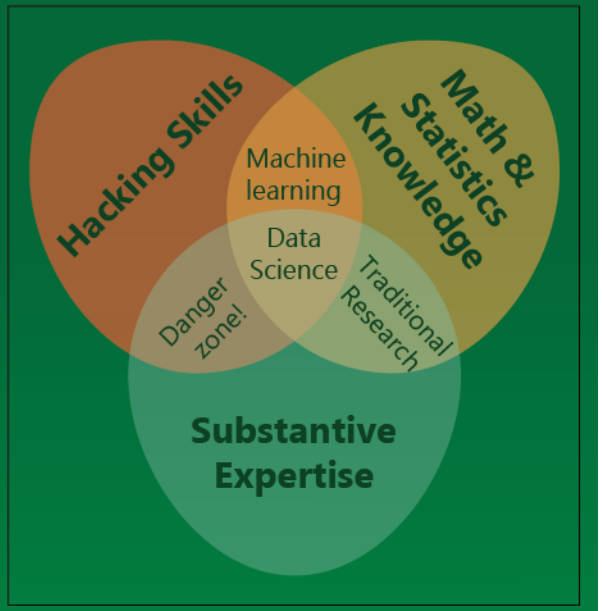
Science of Studying Data :
Programming + Statistics + Business

SKILLS REQUIRED...

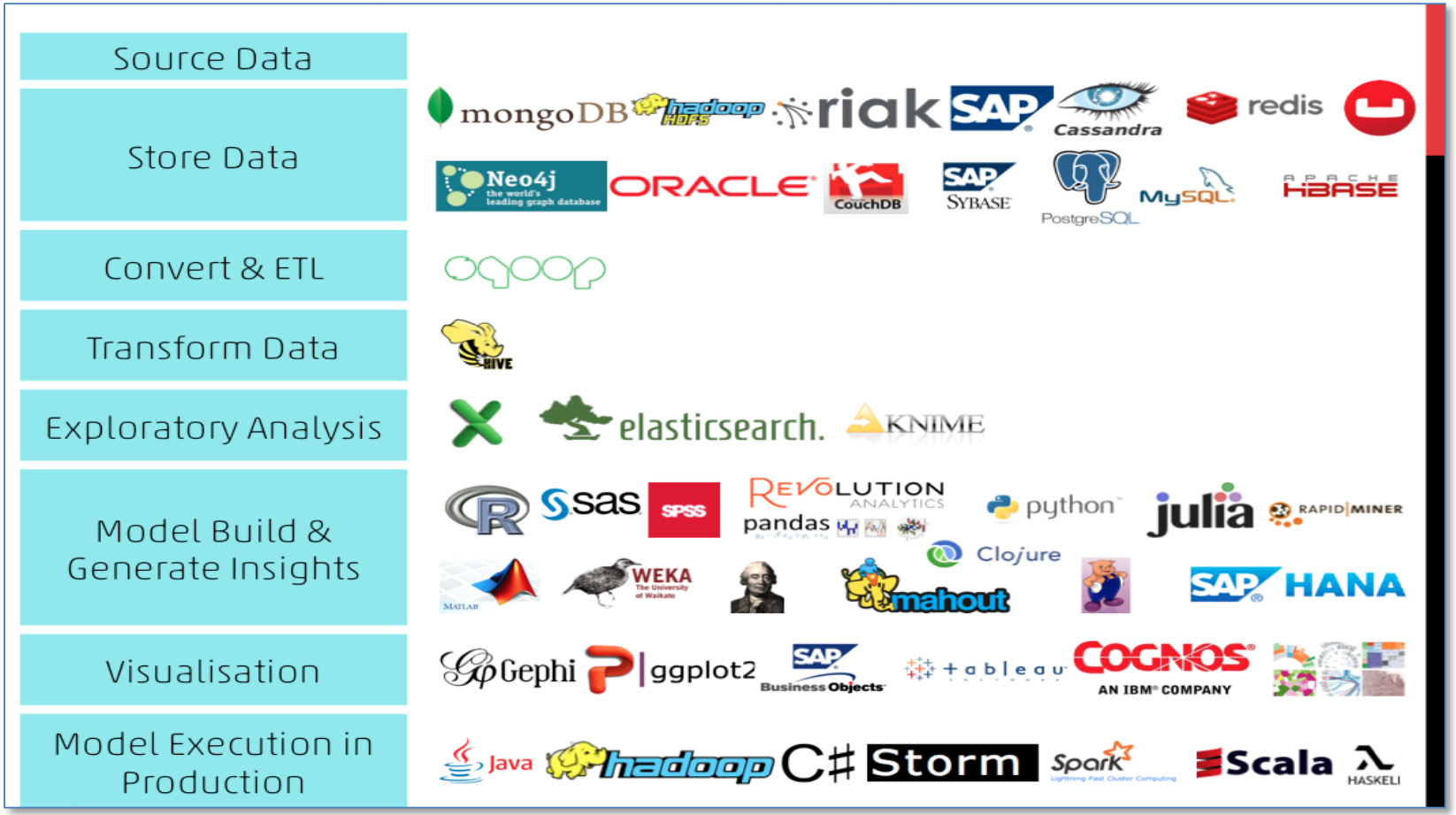




Conway's Diagram



Data Science: Tools and Technologies available



Data Science: Multiple Roles

Data Science job titles :

- Business Intelligence Specialist
- Data Modeler
- Data Curation Specialist
- Metadata Librarian
- Data Mining Specialist
- Market Research Analytics
- SAS Programmer
- Data Mining Specialist
- Data Analytics Engineer
- Data Visualization Specialist
- Digital Curation Librarian
- CRM Analyst
- Data Manager
- Data Journalist

Data Science job Roles:

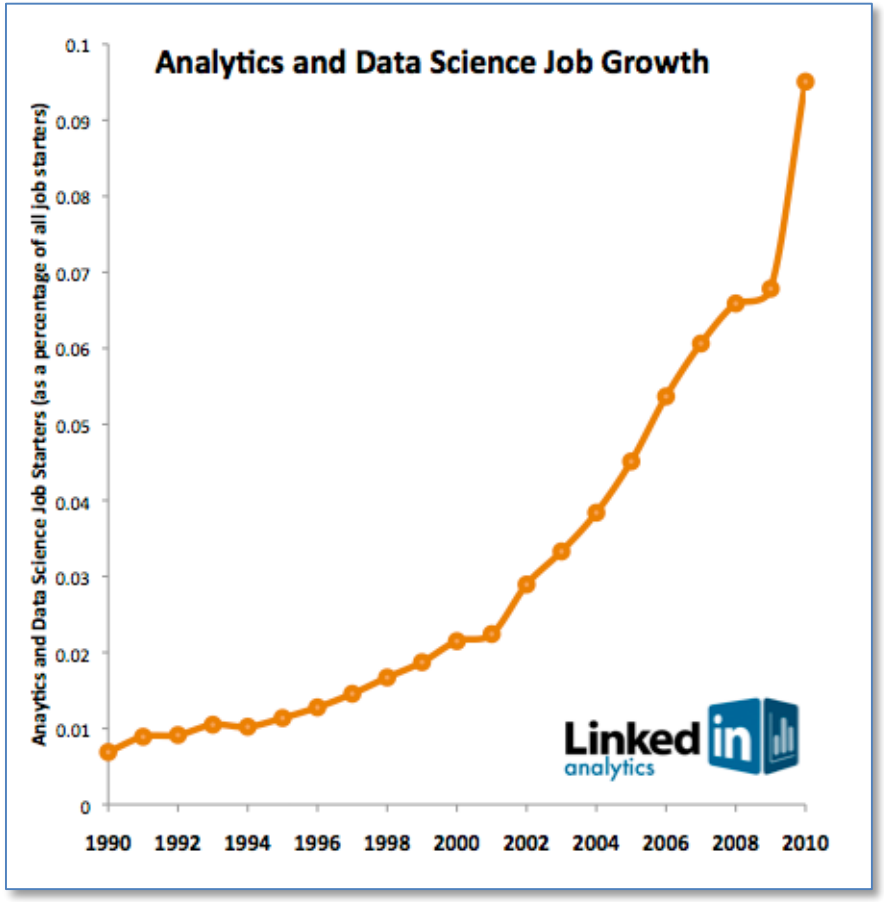
• **Data creator:** Researchers with domain expertise who produce data. These people may have a high level of expertise in handling, manipulating and using data.

• **Data scientist:** People who work where the research is carried out – or in close collaboration with the creators of the data – and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology.

• **Data manager:** Computer scientists, information technologists or information scientists and who take responsibility for computing facilities, storage, continuing access and preservation of data.

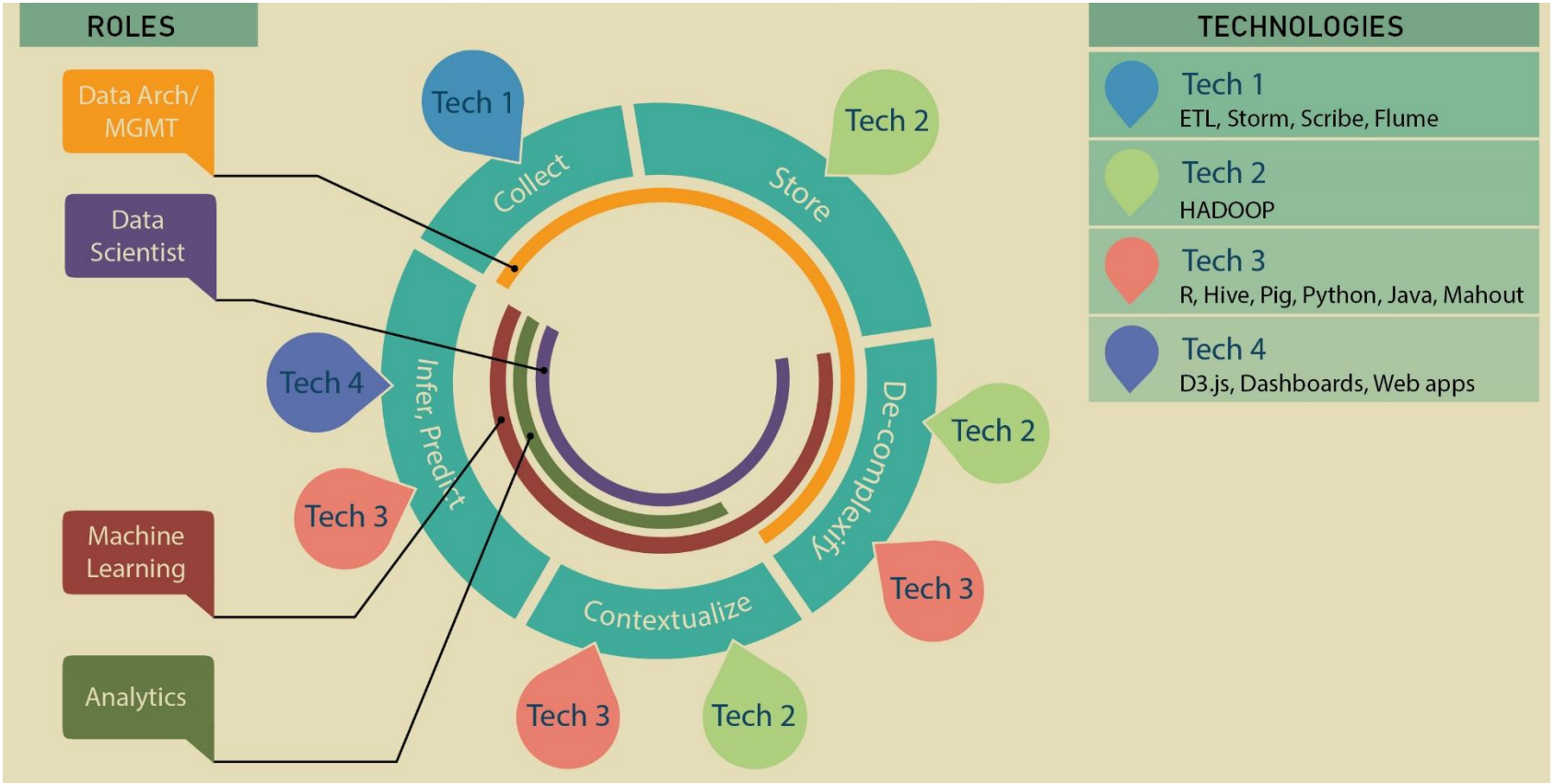
• **Data librarian:** People originating from the library community, trained and specialising in the curation, preservation and archiving of data.

Data Science: Prospects

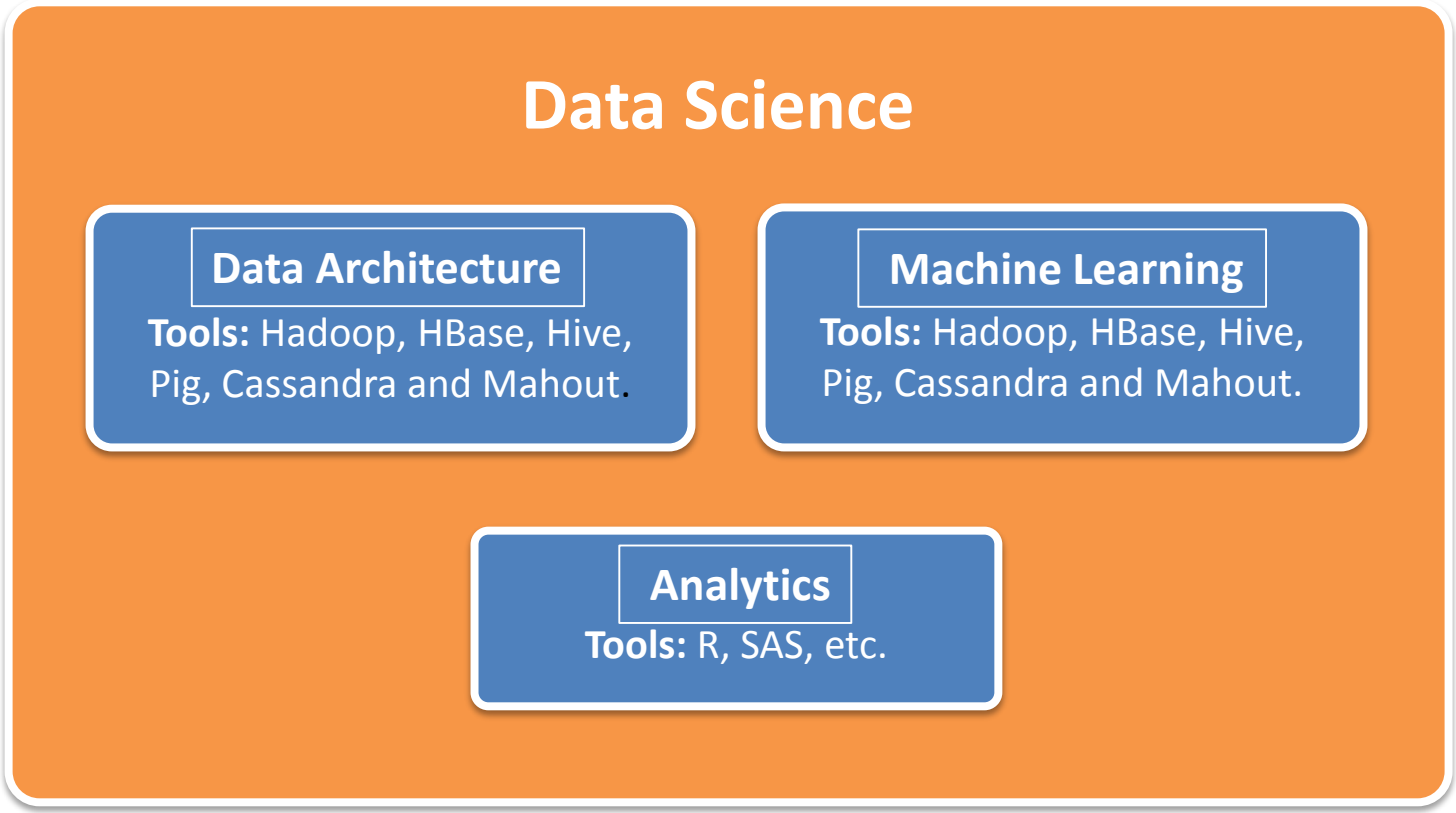


Courtesy LinkedIn Corp.

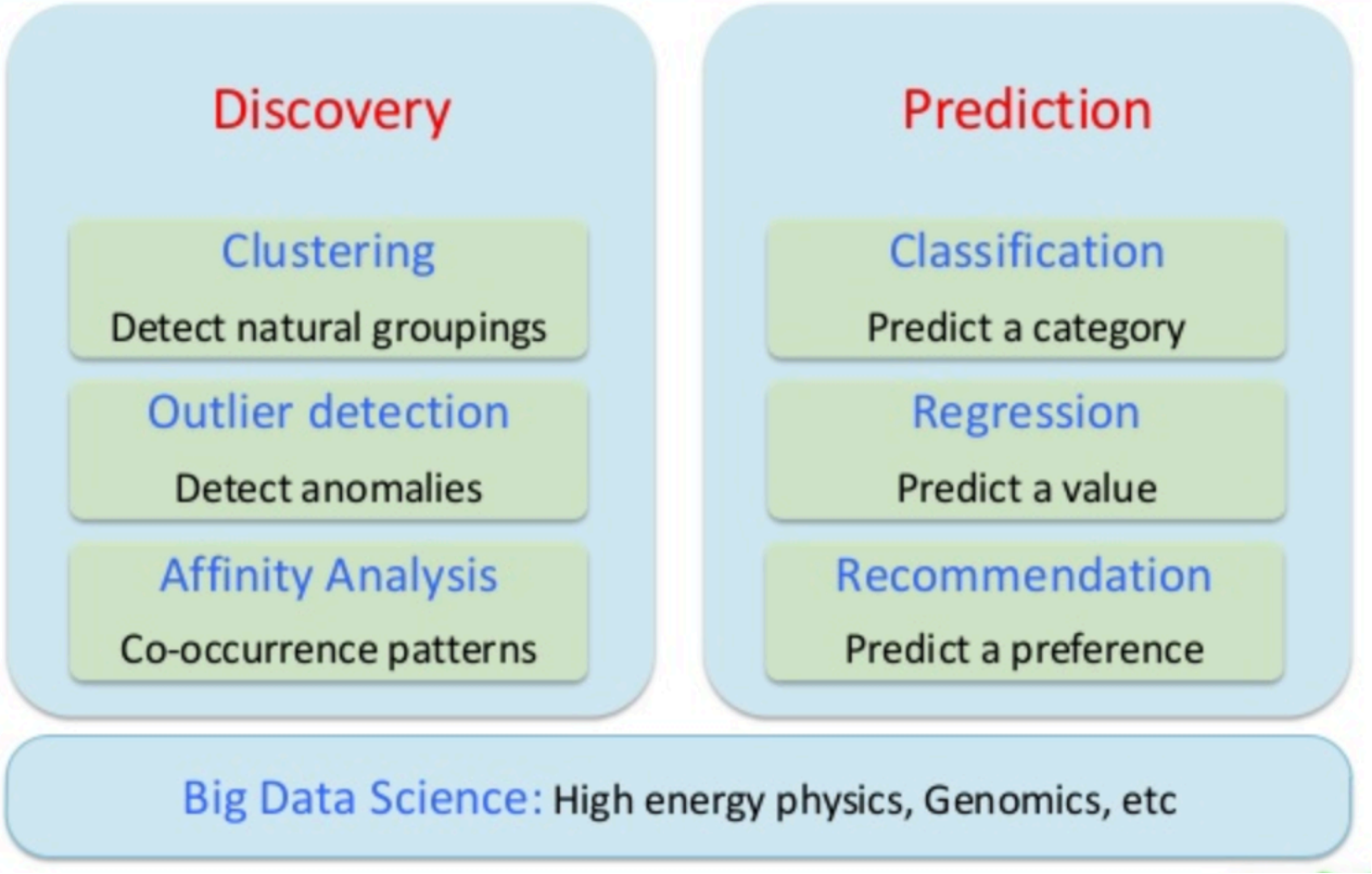
Data Science: What it actually is



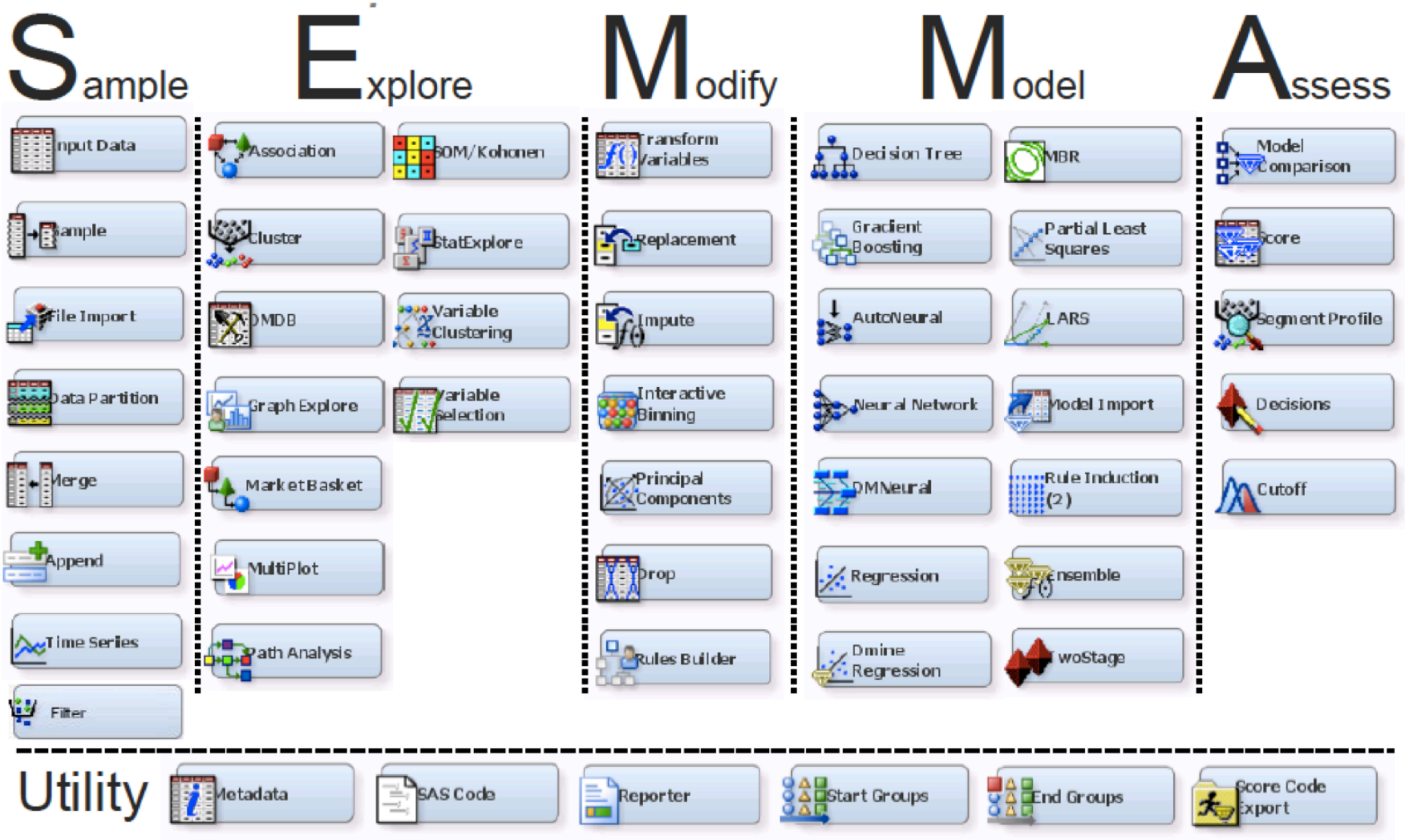
Data Science: Core Components



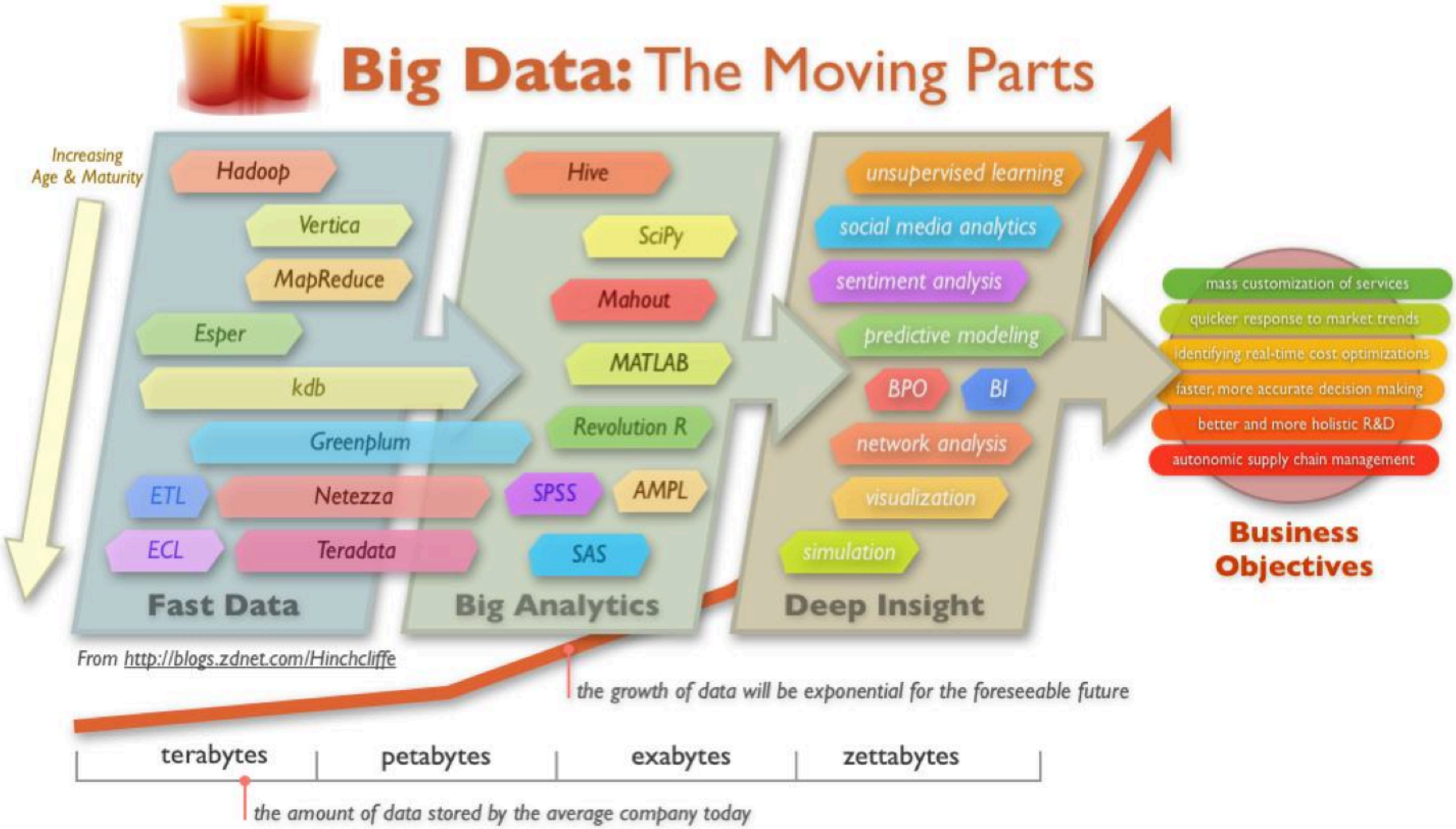
Data Science Tasks



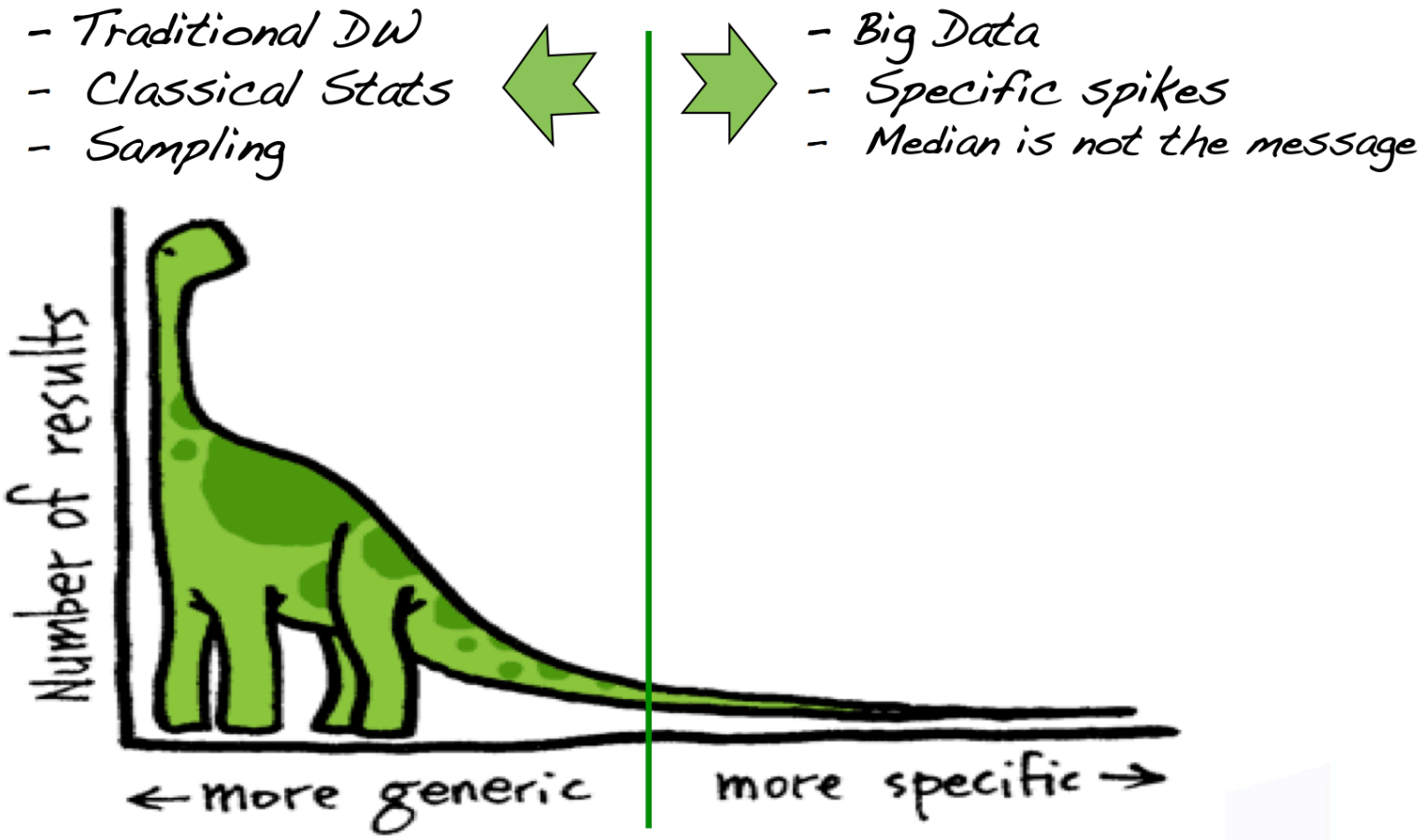
Data Mining Process



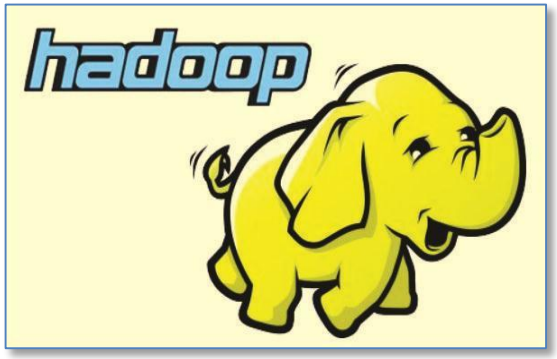
Big Data Technology Vendors



Big Data Science- A perspective



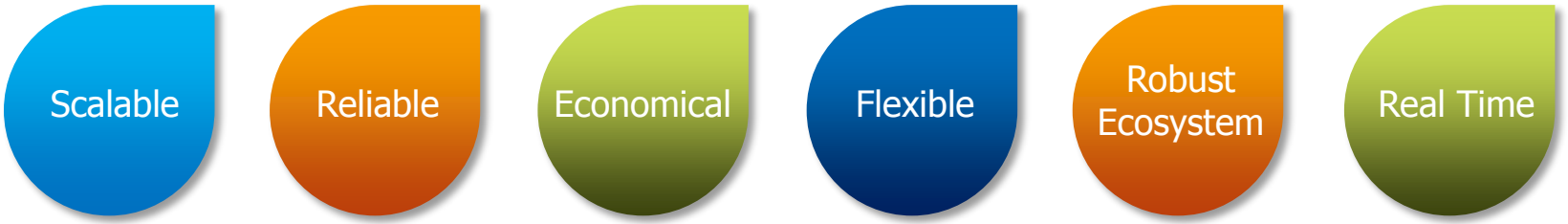
Introduction to Hadoop



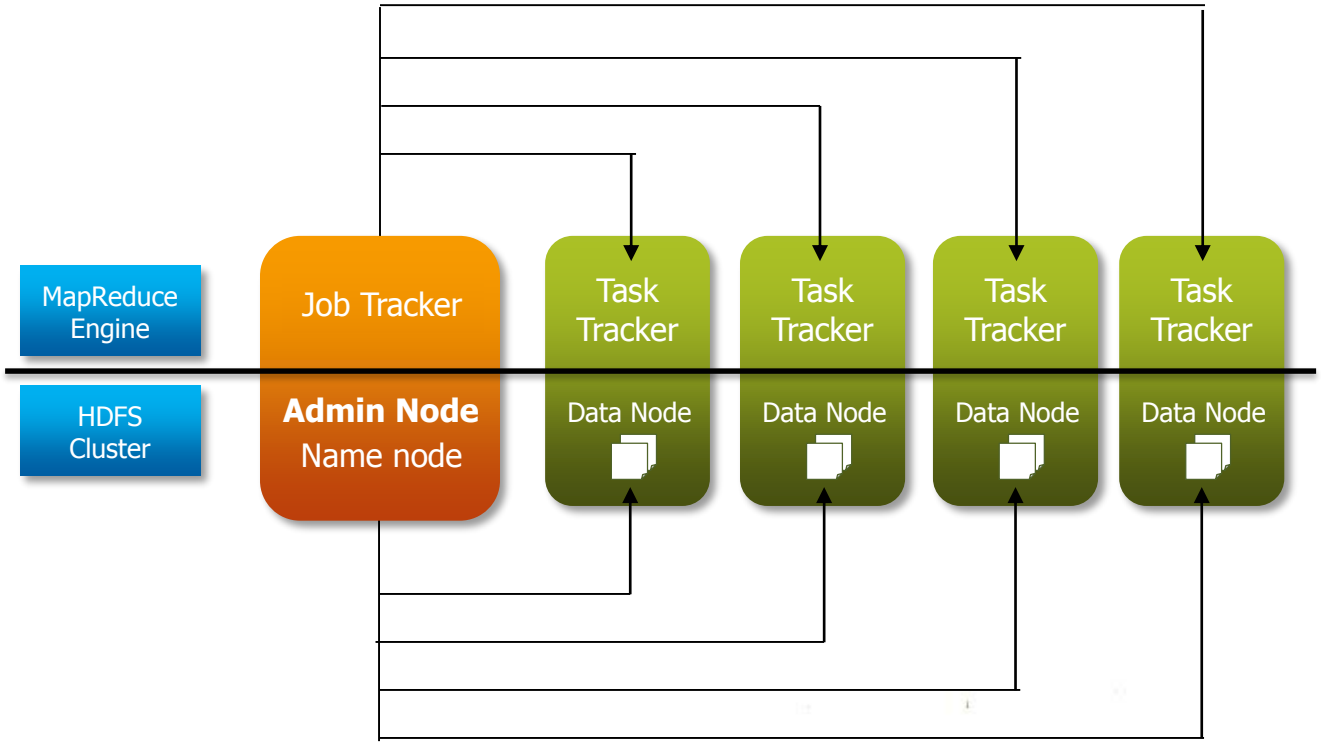
Introduction to Hadoop

- ✓ Apache Hadoop is a **framework** that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model.
- ✓ It is an **Open-source Data Management** with scale-out storage & distributed processing.
- ✓ In 2004, Google published a paper on a process called **MapReduce**.
- ✓ MapReduce framework provides a parallel processing model and associated implementation to process huge amount of data.
- ✓ Therefore, an implementation of MapReduce framework was adopted by an Apache open source project named Hadoop.

Hadoop Key Characteristics

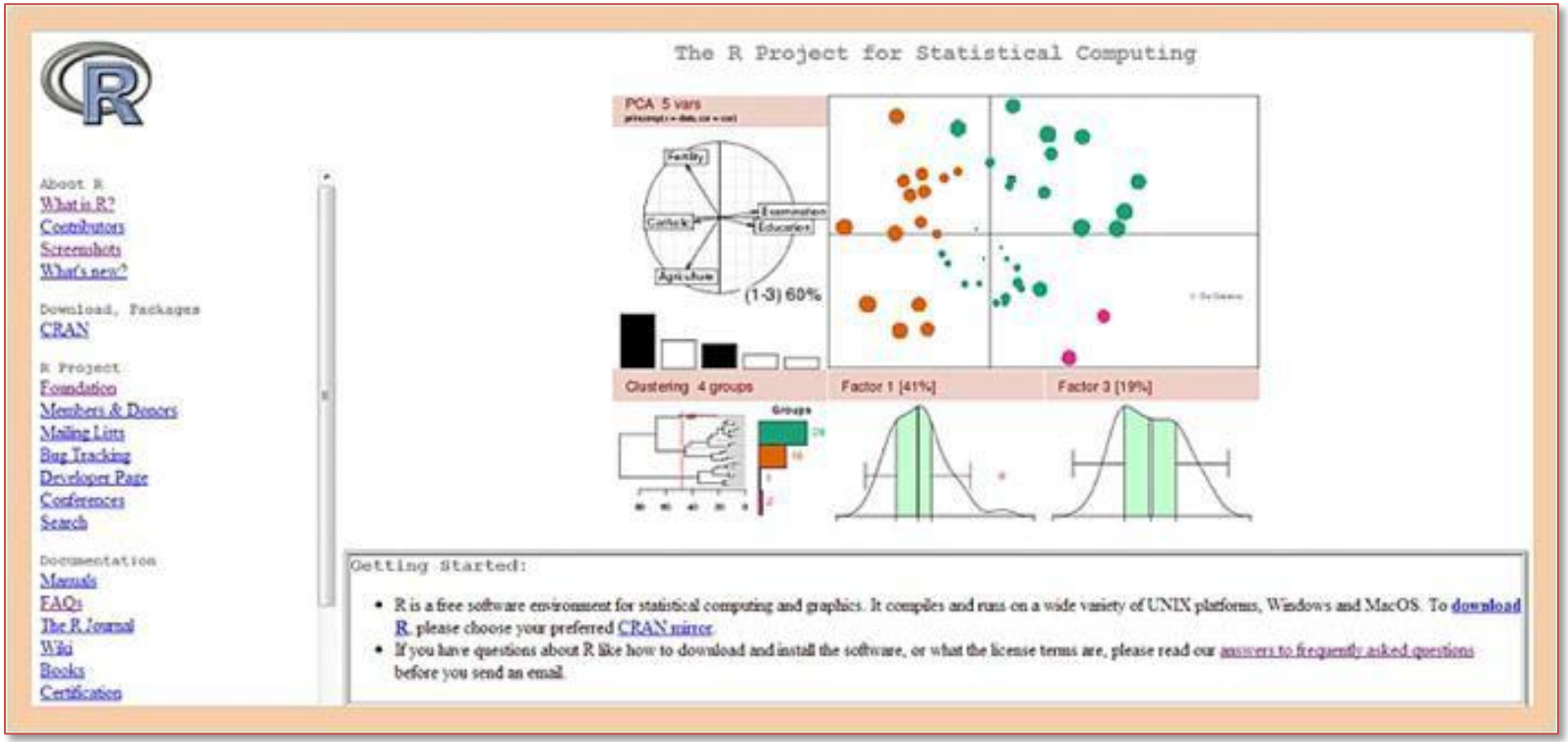


Hadoop Core Components



Analytics with R

Analytics with R



R : Pros and Cons

R is open source and free

R has lots of packages

R can be customized

R has the most advanced graphics

R has GUI to help make learning easier

R can connect to many database and data types

but

R has a steep learning curve

multiple packages and ways to do the same thing

by default stores memory in **RAM**

you need much better programming skills

customization needs command line

you need to know which package to use

Comparing R and Base SAS* /SAS Stat*

R is open source and **free**

Open source R has support from email lists, twitter, stack overflow

R is slower on the desktop than base SAS for datasets ~4-5 gb

R has much much **better graphics**

You can create custom functions in R easily

R has multiple GUI that are **free**

but

Base SAS* , SAS/Stat*, SAS/ET*, SAS/OR*, SAS/Graph* are expensive relatively because of annual licenses

SAS Institute* products have dedicated support and extensive documentation

by default **R** stores memory in **RAM,**so we can use the cloud

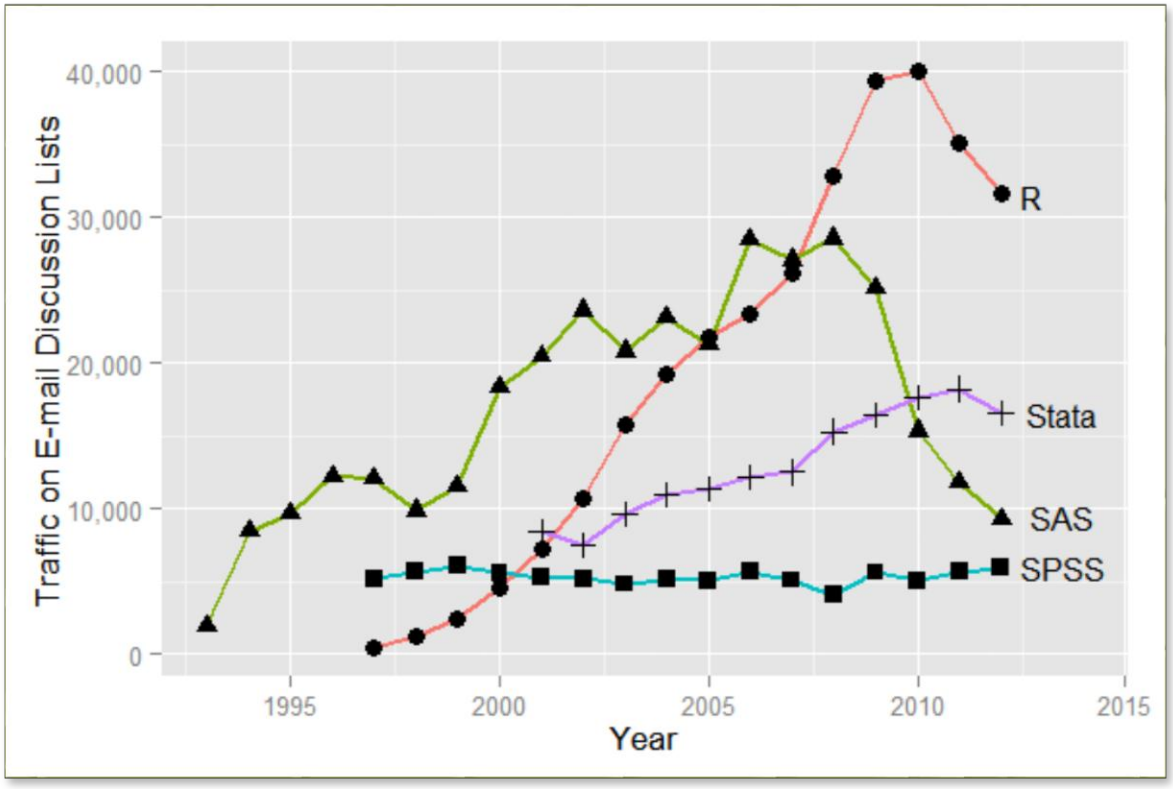
you need much **better** programming skills

Customization needs **command line**

SAS GUI are more expensive

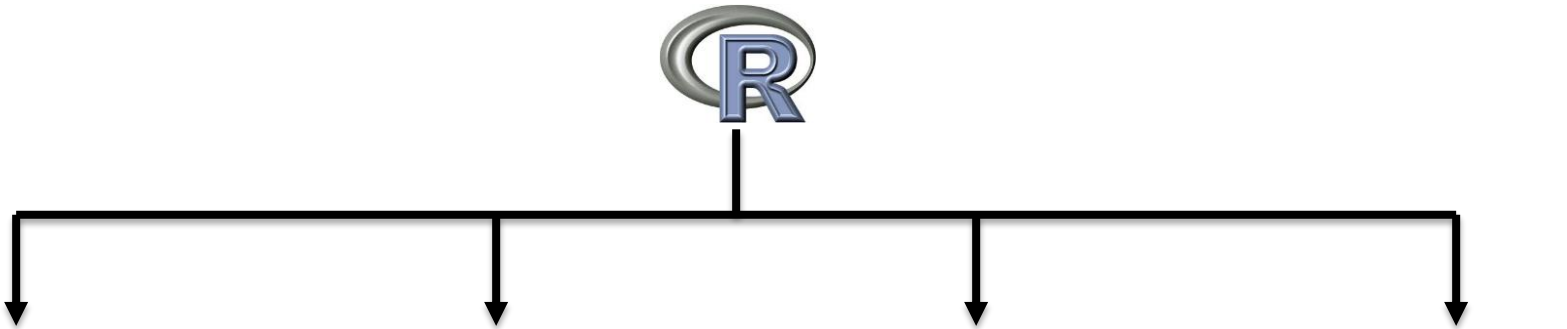
Comparing R and others

http://r4stats.com/articles/popularity/



Introduction to R Programming language

www.r-project.org/about.html



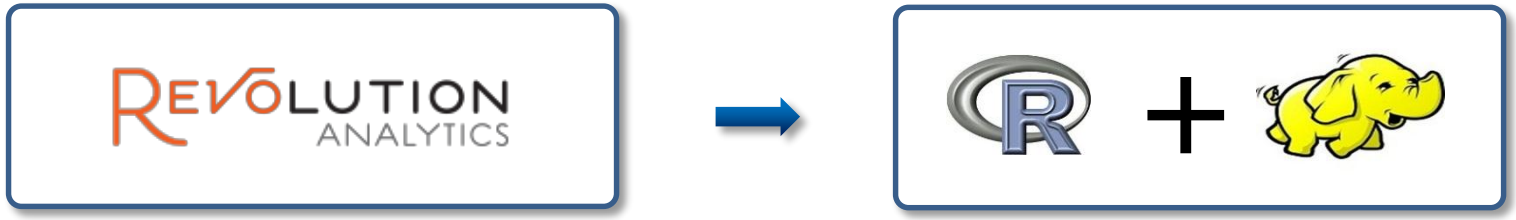
- ✓ History
- ✓ Evolution
- ✓ Current state

- ✓ Open Source
- ✓ Free
- ✓ Widely Recognized

- ✓ Official Website
- ✓ R Core
- ✓ Creators

- ✓ R Journal

R and Hadoop Integration



- R and Hadoop are a natural match in Big Data Analytics and visualization.

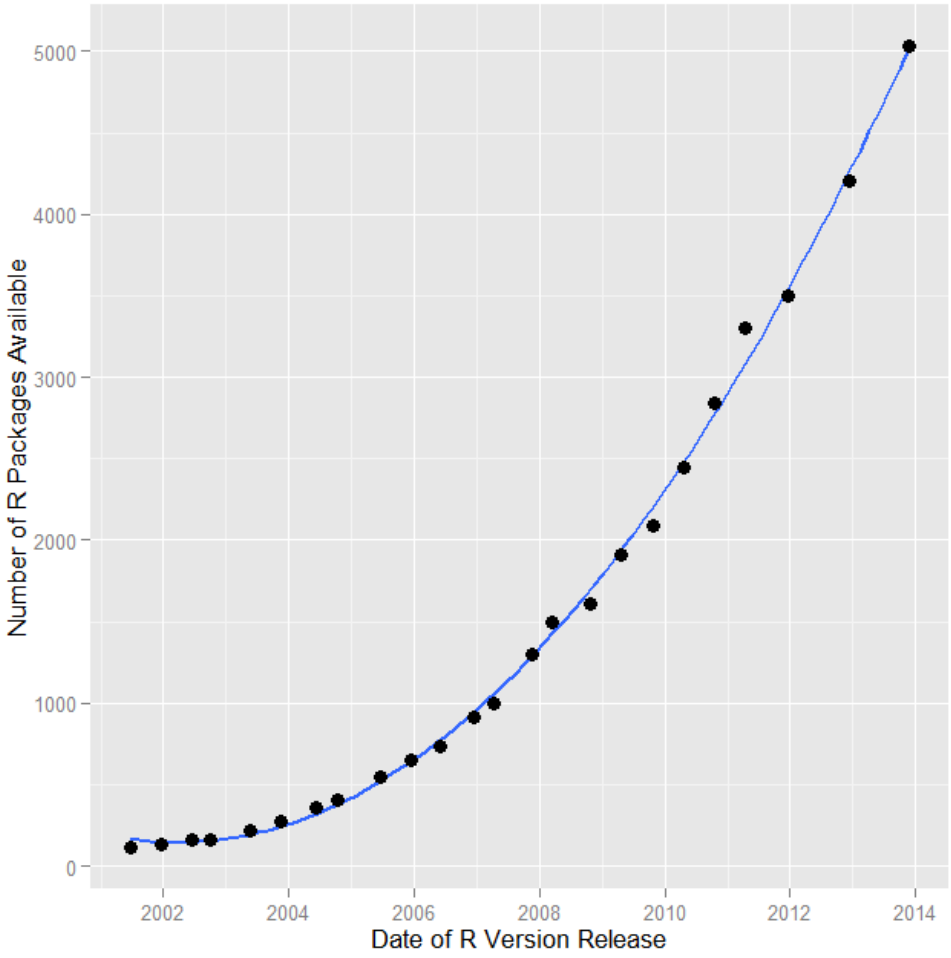
- One of the most well-known R packages to support Hadoop functionalities is : **RHadoop**
- Rhadoop was developed by RevolutionAnalytics.

- RHadoop is a collection of three R packages: rmr, rhdfs and rhbase
- rmr package provides Hadoop MapReduce functionality in R, rhdfs provides HDFS file management in R and rhbase provides HBase database management from within R.

R Intro

R is:

- an open source implementation of S
- a language and an environment
- provides methods for both statistical and graphical data analysis
- runs on Windows, Mac, and Unix systems



What is R?

Source: www.r-project.org

Open-source stat package with visualization

Vibrant community support.

One-line calculations galore!

Steep learning curve but worth it!

Insight into statistical properties and trends…

or for machine learning purposes…

or Big Data to be understood well.

R Open Source Analytics

- Most advanced statistical analysis software available
- Half the cost of commercial alternatives
- 2M+ Users
- 3,000+ Applications

Statistics	Finance
Predictive Analytics	Life Sciences
Data Mining	Manufacturing
Visualization	Retail
	Telecom
	Social Media
	Government

Forbes

Power in the Numbers

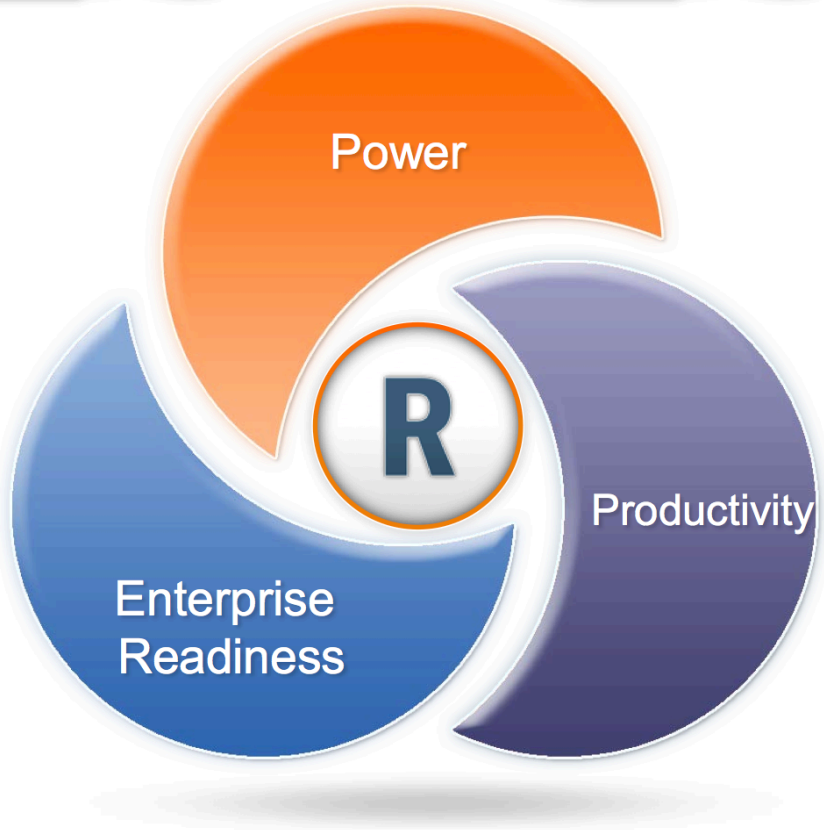
Quentin Hardy, 05.06.10, 09:00 AM EDT

The professor who invented analytic software for the experts now wants to take it to the masses

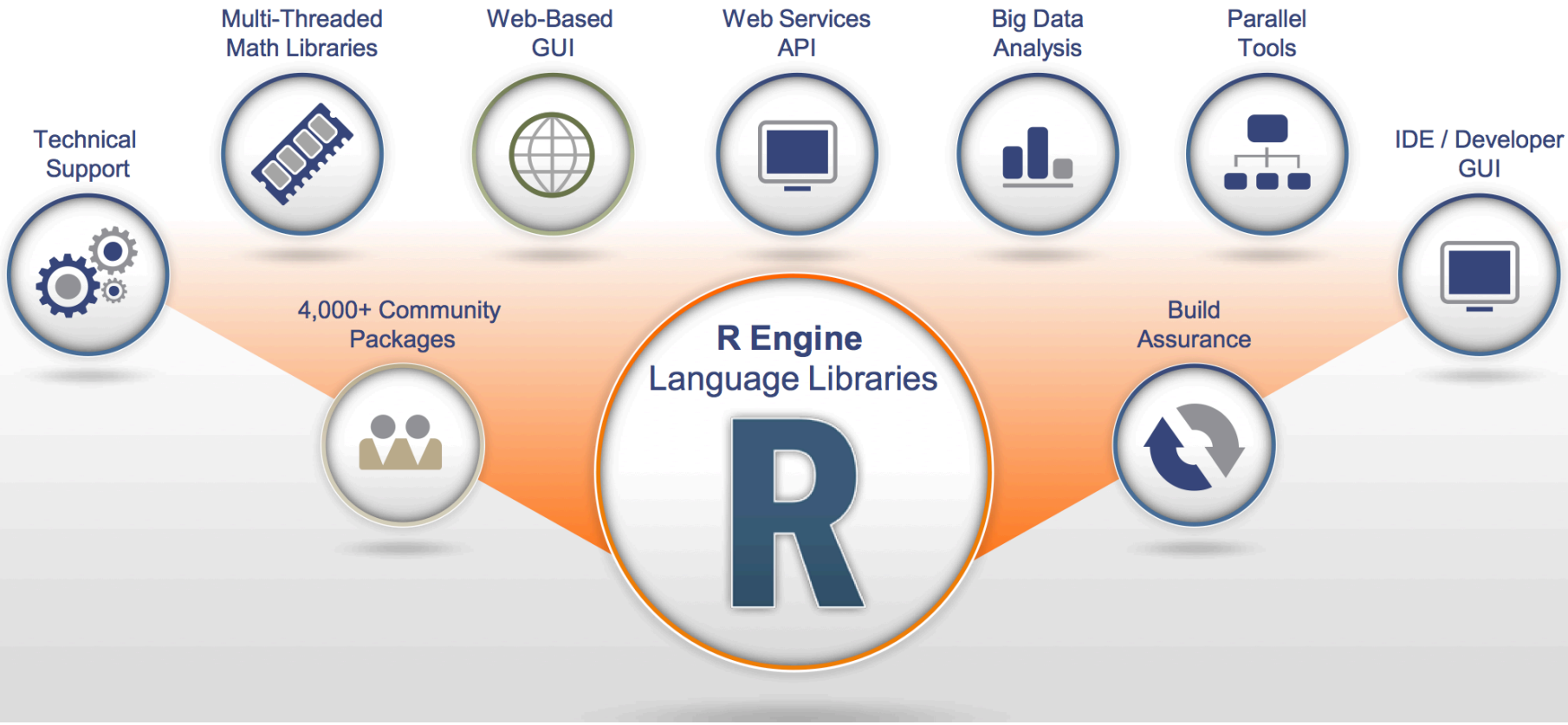
The New York Times

Data Analysts Captivated by R's Power

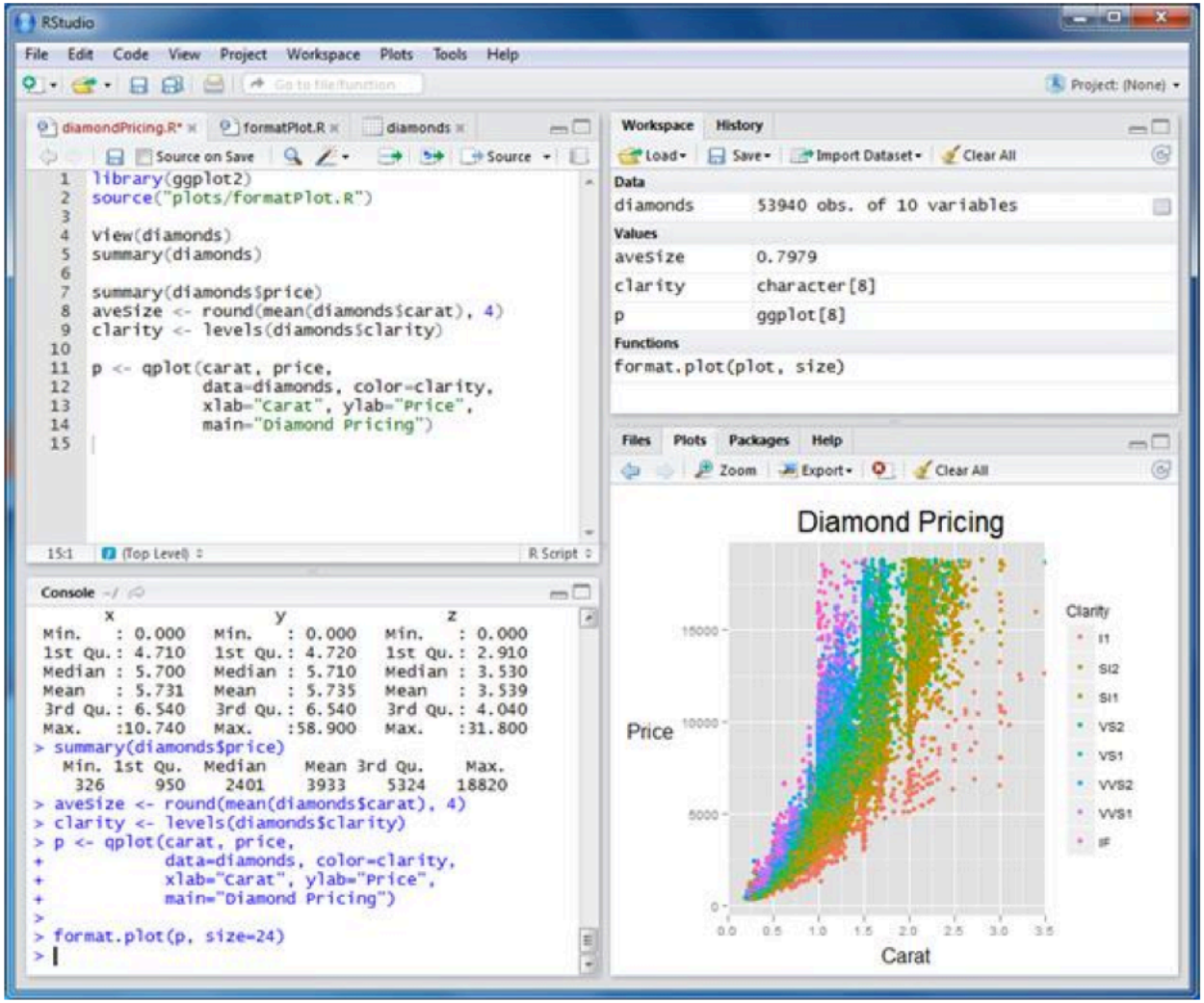
By ASHLEE VANCE
Published: January 6, 2009



R Universe


Technical Support
Multi-Threaded Math Libraries
Web-Based GUI
Web Services API
Big Data Analysis
Parallel Tools
IDE / Developer GUI
4,000+ Community Packages
R Engine
Language Libraries
R
Build Assurance

R-Studio

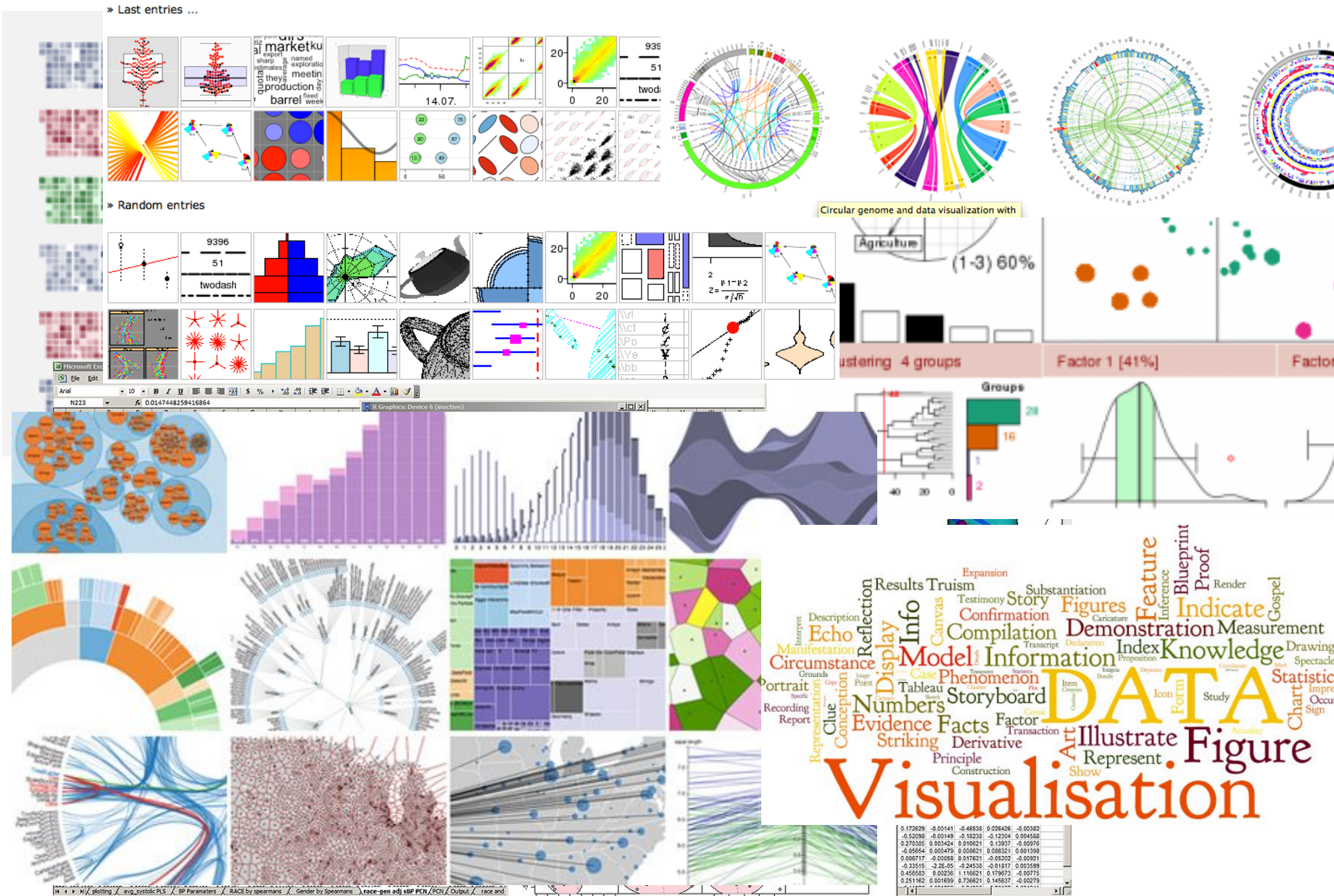


Source: www.rstudio.com/ide/

CASE STUDIES AND DEMO USING R

Source: www.r-project.org

Examples of R Visualizations



Machine Learning

Machine Learning: Mahout

- ✓ Machine Learning is a class of algorithms which is data-driven, i.e. unlike "normal" algorithms it is the data that "tells" what the "good answer" is.

- ✓ **Example:**

  An hypothetical non-machine learning algorithm for face recognition in images would try to define what a face is (round skin-like-colored disk, with dark area where you expect the eyes etc).

  A machine learning algorithm would not have such coded definition, but will "learn-by-examples": you'll show several images of faces and not-faces and a good algorithm will eventually learn and be able to predict whether or not an unseen image is a face.

Mahout Overview

ML is all over the web today

Mahout is about scalable machine learning

MapReduce magic in action

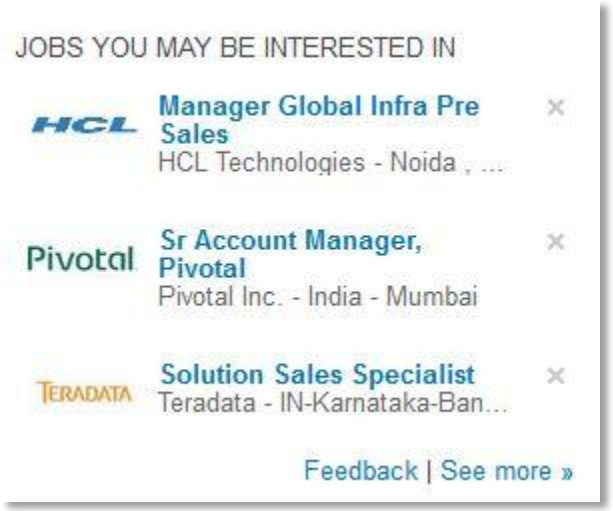
Mahout has functionality for many of today's common machine learning tasks

Machine Learning with Mahout

Write intelligent applications using Apache Mahout

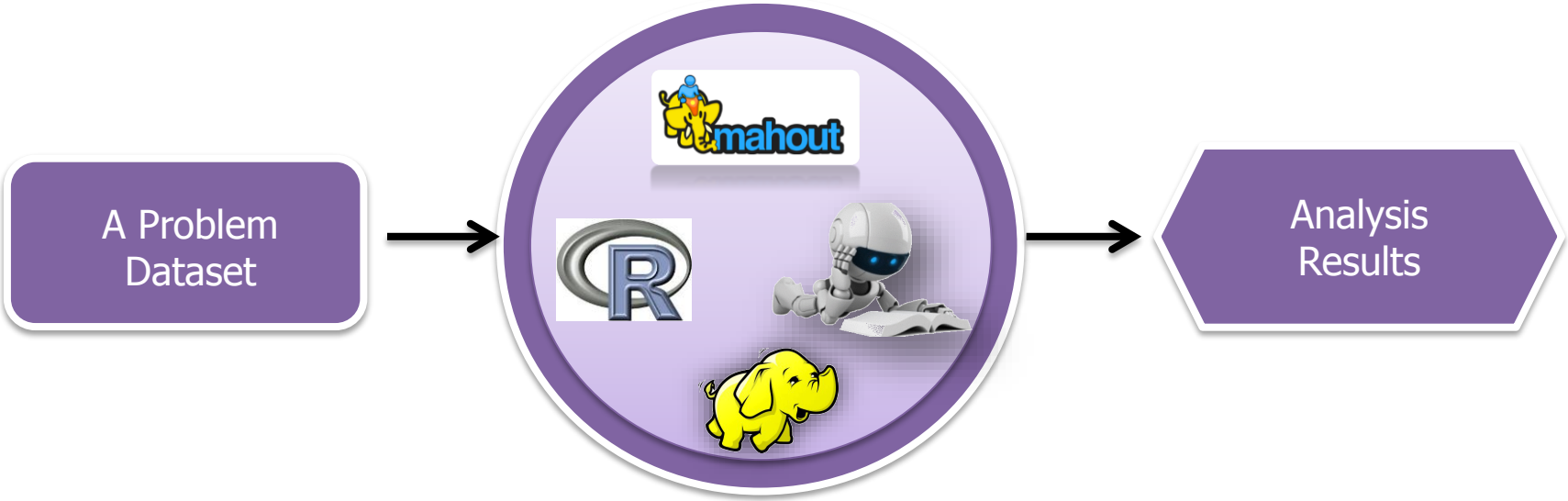
Hadoop and MapReduce magic in action

LinkedIn Recommendations



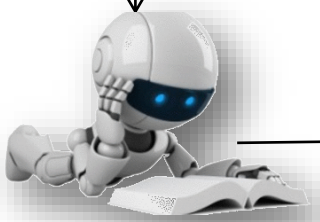
https://cwiki.apache.org/confluence/display/MAHOUT/Powered+By+Mahout

Datasets


A Problem Dataset
mahout
Analysis Results

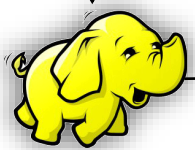
Understanding the Problem Statement and defining the solution

Understanding the Machine Learning algorithm to be used



Implementing Machine Learning algorithm in R on the smaller dataset.

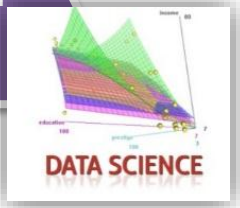
Implementing Machine Learning in Hadoop on Big Data.



Exploring ways to integrate R with Hadoop

REVOLUTION ANALYTICS

Visualisation of the analysis



Dataset 1

Domain of the Dataset:

Communications and Media. However, the application of the algorithm is not limited to only Communications and Media. The technique is useful for any domain which requires organizing documents to improve retrieval and support browsing.

Problem Statement:

Clustering / Grouping documents based on their contents.
A top media company wants to browse through the popular news from a collection that appeared on the Reuters newswire in 1987. Organizing the documents in coherent categories will be very useful for systematic browsing of the document collection.

R Implementation

First we will understand the implementation of the technique in R on a smaller dataset

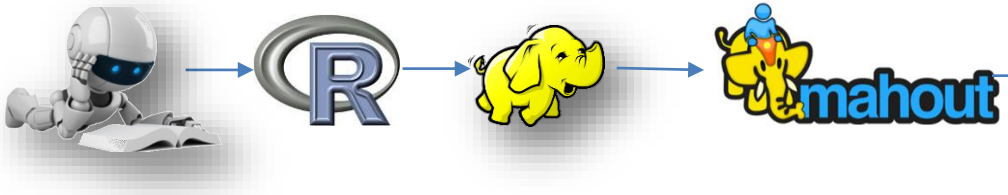
Hadoop Implementation

Then we will understand how to achieve document clustering on Big Data using Mahout libraries on Hadoop

Machine Learning Implementation

K-Means Clustering can be implemented on this dataset

Communications and Media Dataset to be Clustered based on their contents



Content-wise Clustered/Grouped documents.

Dataset 2

Domain of the Dataset:

Products and Retail. However, the application of the algorithm is not limited to only Products and Retail. The technique can be applied wherever we want to discover the co-occurrence relationship amongst various activities.

Problem Statement:

Market Basket Analysis.
A retail outlet wants understand the purchase behavior of a buyer. This information will enable the retailer to understand the buyer's needs and rewrite the store's layout accordingly, develop cross-promotional programs, or even capture new buyers. The analysis might tell a retailer that customers often purchase shampoo and conditioner together, so putting both items on promotion at the same time would not create a significant increase in profit, while a promotion involving just one of the items would likely drive sales of the other.

R Implementation

Understand the implementation of the technique on a smaller dataset

Hadoop Implementation

Understand how to achieve the same on Big Data using Mahout libraries on Hadoop

Machine Learning Implementation

The technique used is Affinity Analysis or Association Rule Mining

Product and Retail Dataset

Market-Basket Analysis

Dataset 3

Domain of the Dataset:

Life Science and Health Care. However, the application of the algorithm is not limited to only Life Science and Health Care . The technique can be applied wherever we want to forecast the occurrence of a event on the basis of certain conditions.

Problem Statement:

A health care organization wants to forecast the onset of diabetes mellitus in Indians using certain set of attributes of patients as input.

R Implementation

Understand the basic implementation of the technique on a smaller dataset using R

Achieve parallel processing on the same algorithm using a parallel processing library provided by Revolution R.

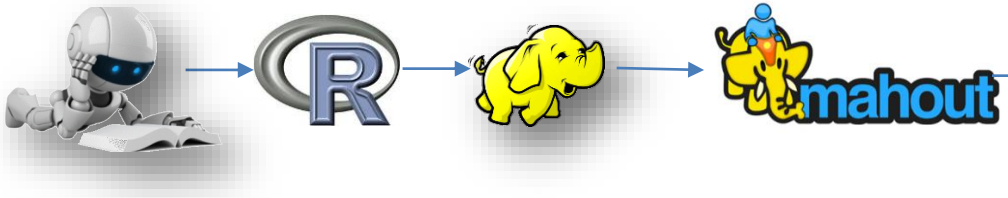
Hadoop Implementation

Understand how to achieve the same on Big Data using Mahout libraries on Hadoop

Machine Learning Implementation

The technique used is Affinity Analysis or Association Rule Mining.

Life Science and Health Care Dataset with some attributes of patients as input.



Forecast the onset of diabetes mellitus in Indians.

Dataset 4

Domain of the Dataset:

Social Media. However, the application of the algorithm is not limited to only Social Media. The technique can be applied wherever we want to put documents into category without going through the contents of all the documents.

Problem Statement:

A Social Media research firm wants to know the trends of topics discussed on Twitter. For easy analysis it wants to classify them in the following categories:

- apparel (clothes, shoes, watches, …)
- art (Book, DVD, Music, …)
- camera
- event (travel, concert, …)
- health (beauty, spa, …)
- home (kitchen, furniture, garden, …)
- tech (computer, laptop, tablet, …)

R Implementation

Understand the basic implementation of the technique on a smaller dataset using R.

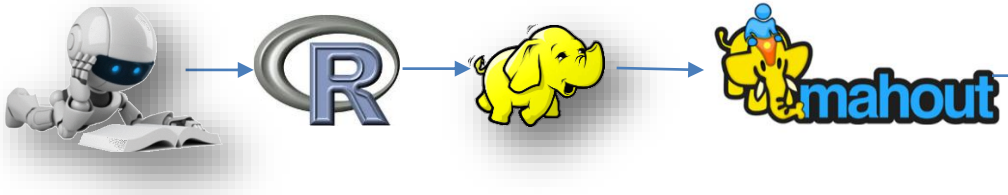
Hadoop Implementation

Understand how to achieve the same on Big Data using Mahout libraries on Hadoop.

Machine Learning Implementation

The technique used is Naïve Bayes Classifier.

Social Media dataset



Categorical Classification of the tweets.