



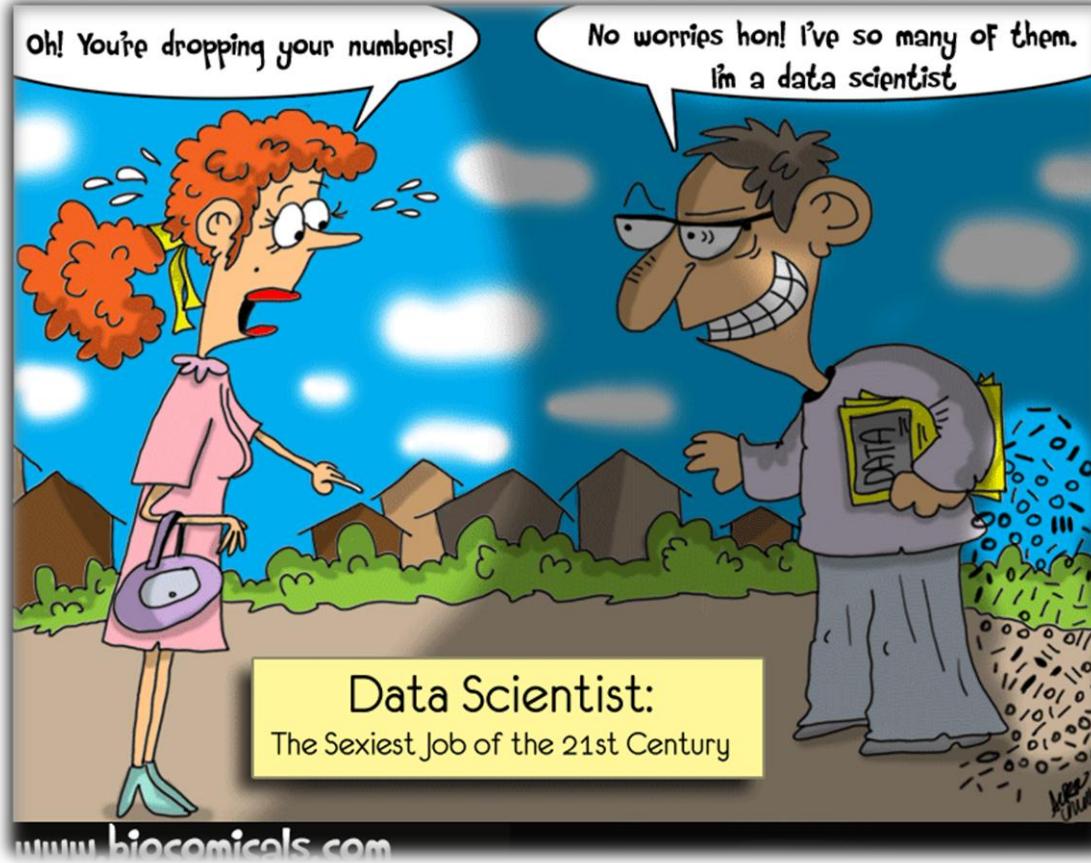
DataInquest

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.



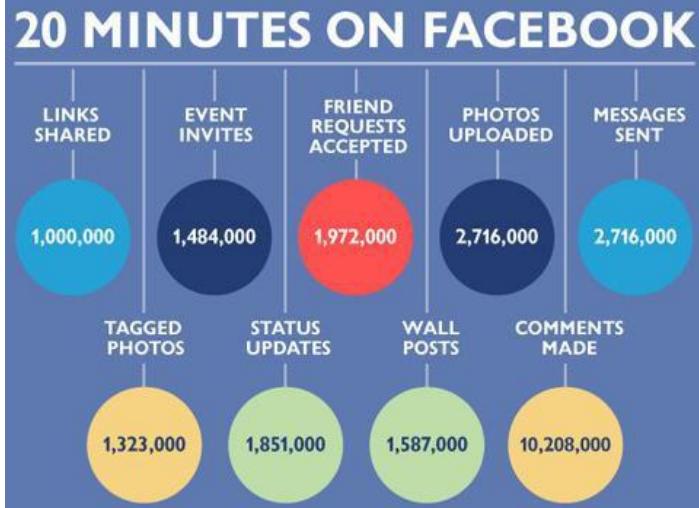
A airline jet collects 10 terabytes of sensor data for every 30 minutes of flying time.



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

Facebook Example

AS OF 2011, THERE ARE 500,000,000 ACTIVE FACEBOOK USERS.
APROX. 1 IN EVERY 13 PEOPLE ON EARTH.
HALF OF THEM ARE LOGGED IN ON ANY GIVEN DAY.



A RECORD-BREAKING
750 MILLION PHOTOS
WERE UPLOADED TO FACEBOOK
OVER NEW YEAR'S WEEKEND.

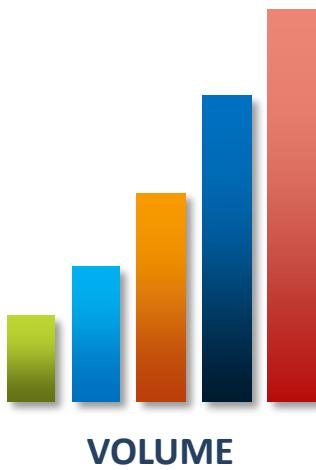
THERE ARE 206.2 MILLION
INTERNET USERS IN THE U.S.
THAT MEANS 71.2% OF THE U.S. WEB AUDIENCE
IS ON FACEBOOK.

- ✓ 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

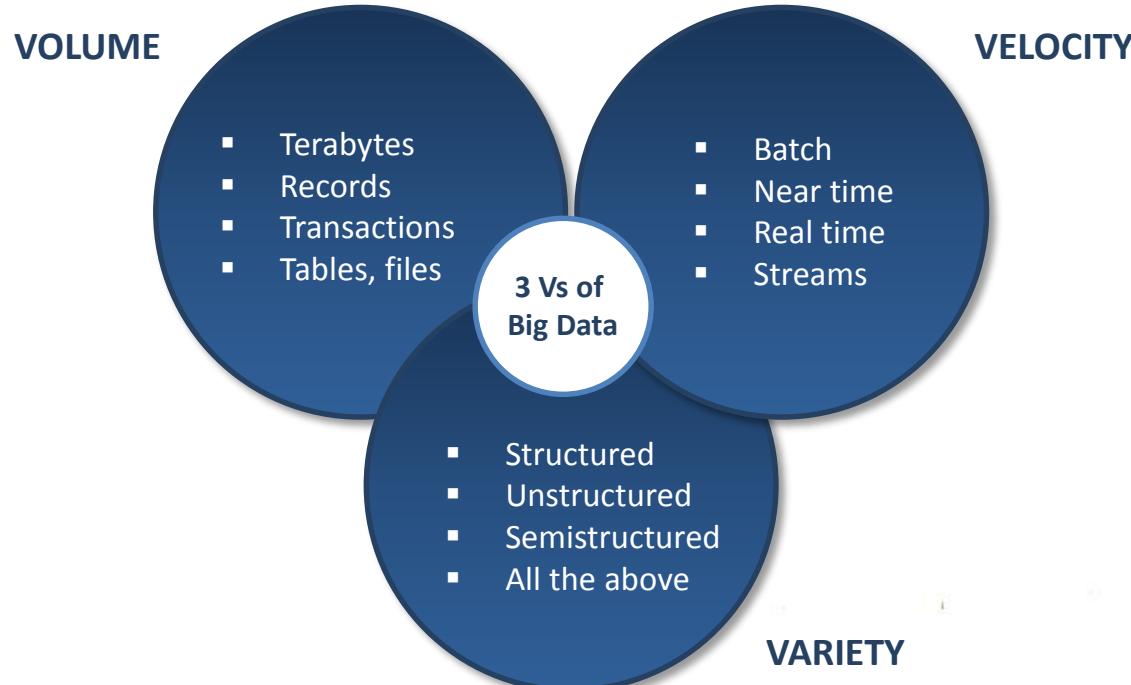
<http://www-01.ibm.com/software/data/bigdata/>



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



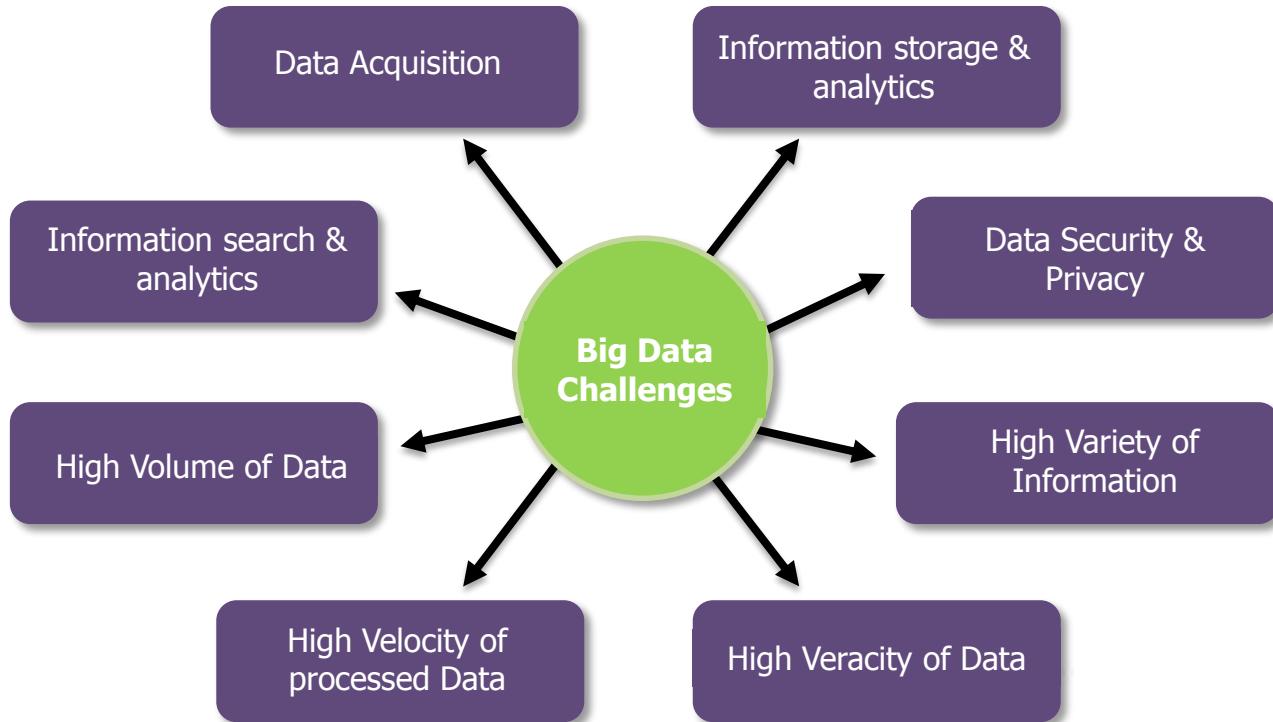
✓ Retail

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



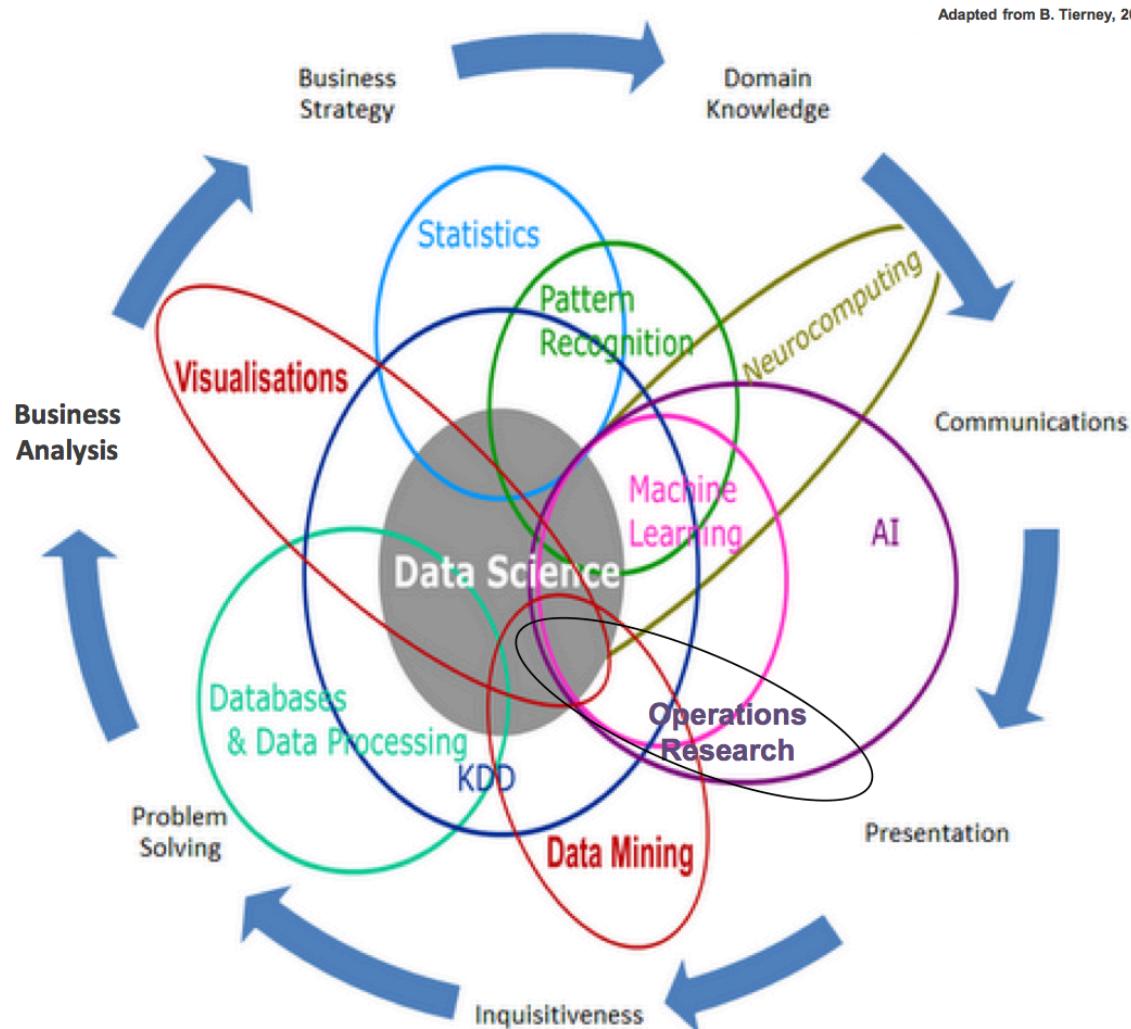
<http://wiki.apache.org/hadoop/PoweredBy>

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.



The screenshot shows the IMDb movie page for "The Wolf of Wall Street" (2013). At the top, there's a navigation bar with links for "Movies, TV & Showtimes", "Celebs, Events & Photos", "News & Community", and "Watchlist". Below the navigation is a search bar with the placeholder "Find Movies, TV shows, Celebrities and more...". The main content area features a large thumbnail image of the movie poster, which shows Leonardo DiCaprio as Jordan Belfort in a suit, standing in front of a crowd of people. To the right of the poster, the movie title "The Wolf of Wall Street" is displayed in bold black text, followed by the year "(2013)". A small yellow square icon with the number "1" indicates it's the first movie in a list. Below the title, the runtime "180 min" and genres "Biography | Comedy | Crime" are listed, along with the release date "25 December 2013 (USA)". A yellow star rating box shows a rating of "8.7". Below the rating, the plot summary reads: "Based on the true story of Jordan Belfort, from his rise to a wealthy stockbroker living the high life to his fall involving crime, corruption and the federal government." Further down, the director "Martin Scorsese" and writers "Terence Winter (screenplay), Jordan Belfort (book)" are mentioned. The stars listed are Leonardo DiCaprio, Jonah Hill, Margot Robbie, and others. At the bottom of the page, there are buttons for "+ Watchlist", "Watch Trailer", and "Share...". A yellow banner at the very bottom states "Top 250 #69 | Nominated for 2 Golden Globes. Another 13 wins & 40 nominations. See more awards »".



This screenshot shows a section of the IMDb website titled "People who liked this also liked...". It displays a grid of movie posters for films like "Lawless" (2012), "CITY OF GOD", "THE SHAWSHANK REDEMPTION", "AMERICAN GANGSTER", and "Trainspotting". To the right of the grid, a detailed movie card for "Lawless" is shown. The card includes the title "Lawless" (2012), the director "John Hillcoat", and the stars "Tom Hardy, Shia LaBeouf, G...". It also shows a yellow star rating box with a rating of "7.3/10". Below the card, there are buttons for "Add to Watchlist", "Next »", and "Prev 6 ▶".

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

The GOOD news:



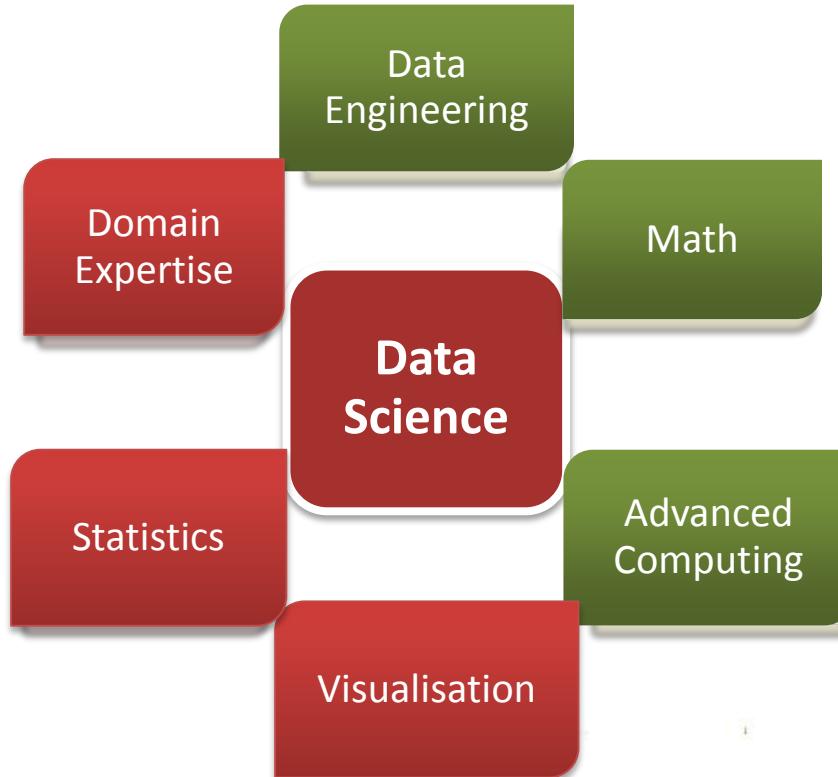
Big Data is here.

The BAD news:



We are struggling to store and analyze it.

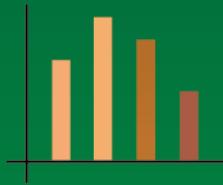
Data Science: Components



WHAT IS DATA SCIENCE ?

Science of Studying Data :
Programming + Statistics + Business

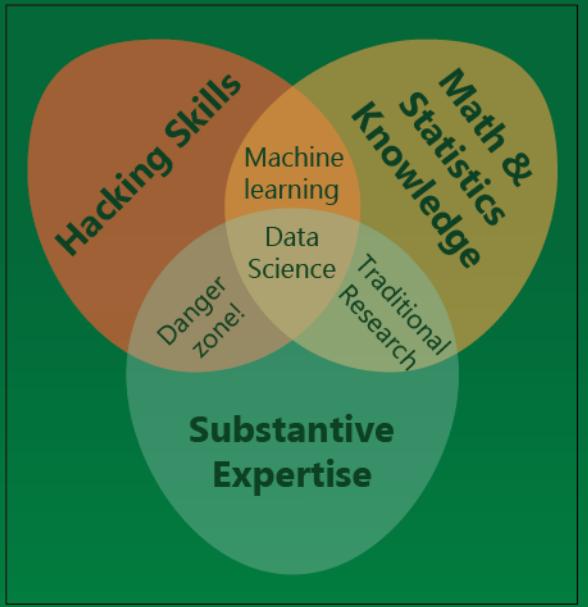
SKILLS REQUIRED...



$$\sqrt{z} \ f \%$$



Conway's Diagram



Data Science: Tools and Technologies available

Source Data



Store Data



Convert & ETL



Transform Data



Exploratory Analysis



Model Build & Generate Insights



Visualisation



Model Execution in Production



Data Science: Multiple Roles

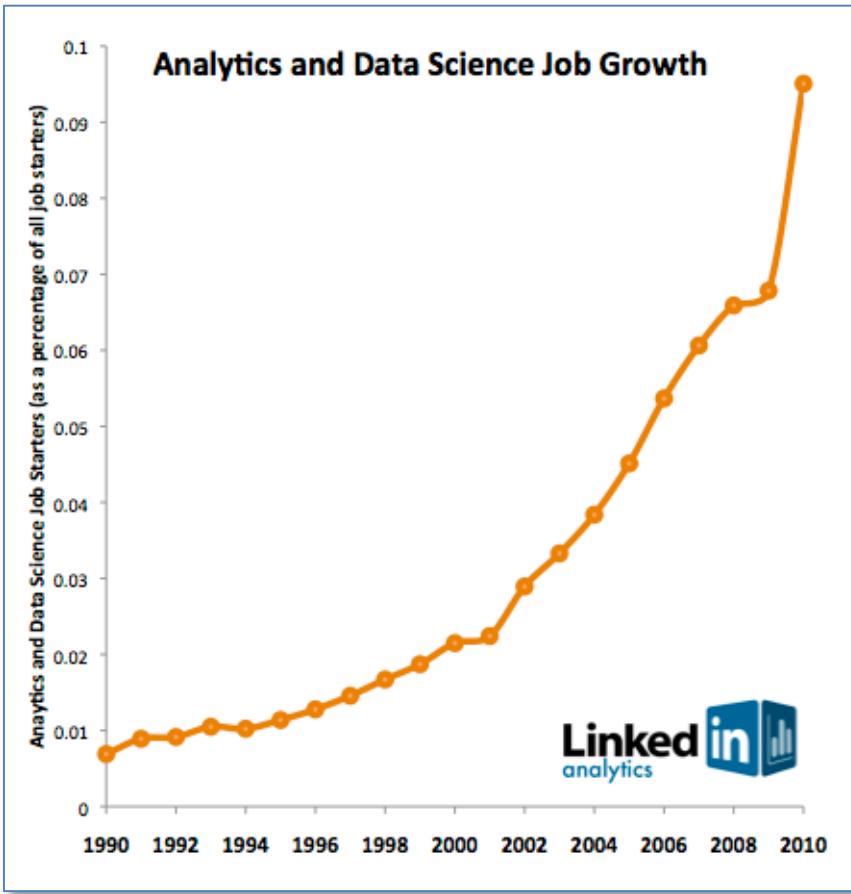
Data Science job titles :

- 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

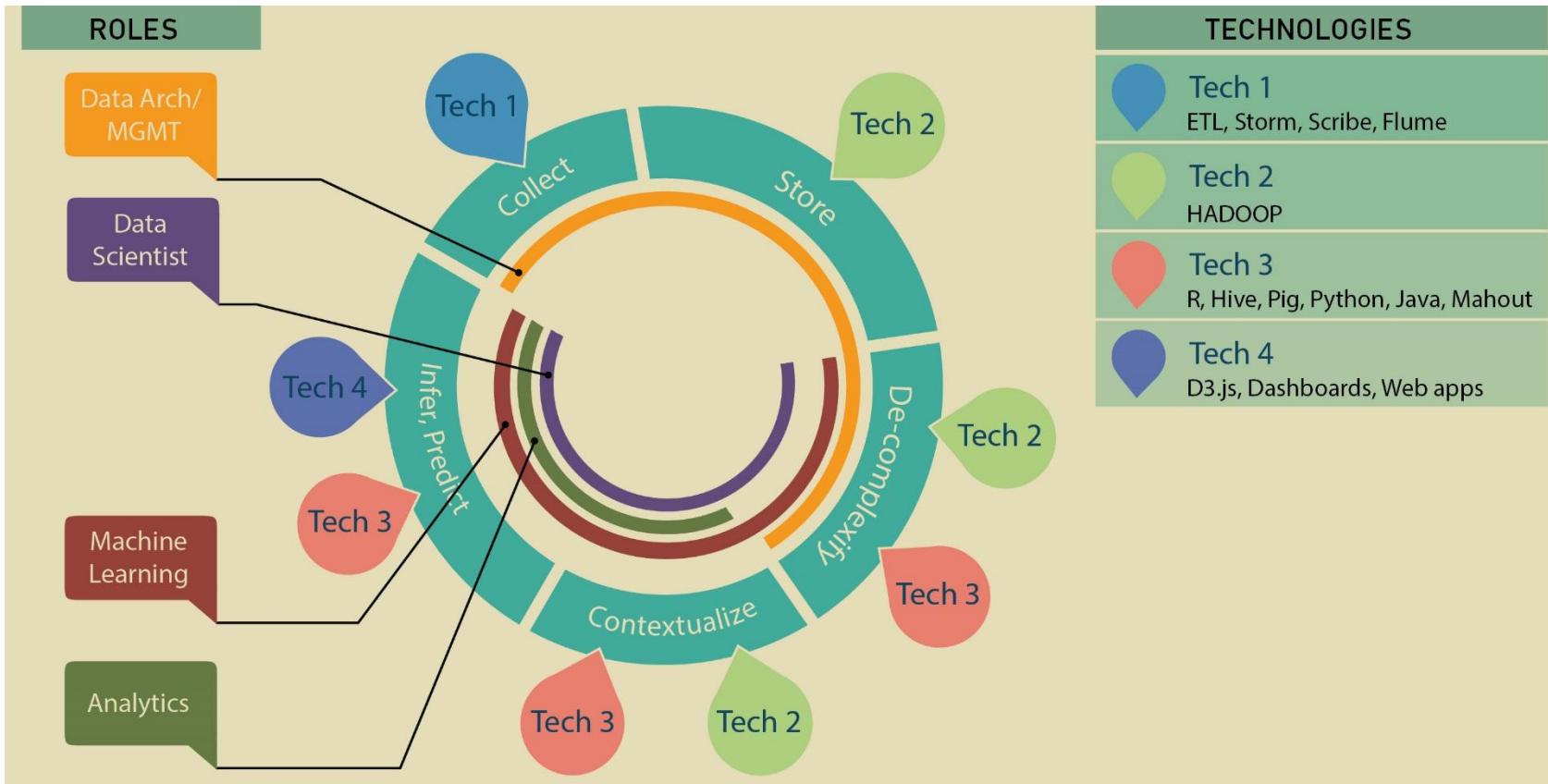
Data Science job Roles:

- **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



Data Science: What it actually is



Data Science

Data Architecture

Tools: Hadoop, HBase, Hive, Pig, Cassandra and Mahout.

Machine Learning

Tools: Hadoop, HBase, Hive, Pig, Cassandra and Mahout.

Analytics

Tools: R, SAS, etc.

Data Science Tasks

Discovery

Clustering

Detect natural groupings

Outlier detection

Detect anomalies

Affinity Analysis

Co-occurrence patterns

Prediction

Classification

Predict a category

Regression

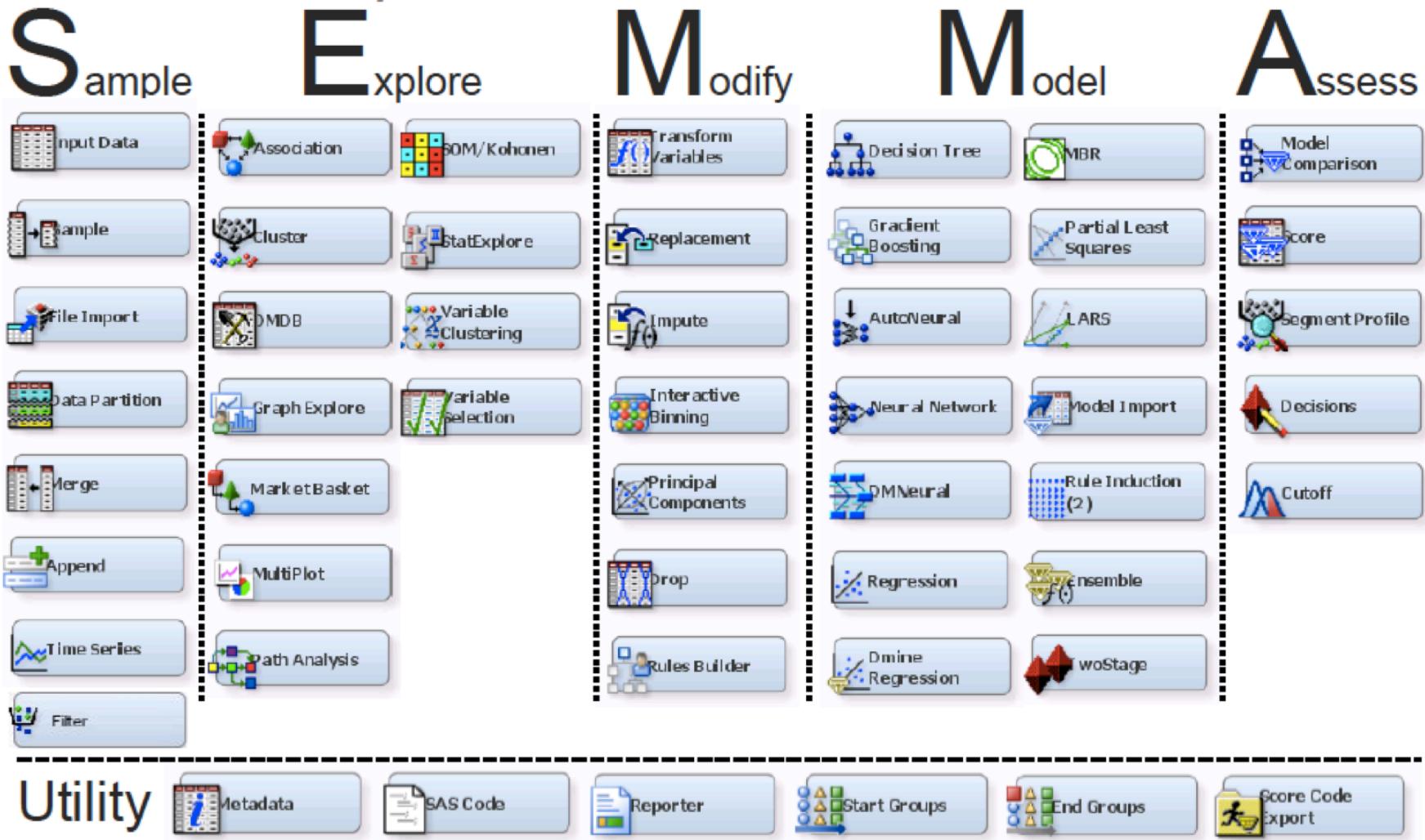
Predict a value

Recommendation

Predict a preference

Big Data Science: High energy physics, Genomics, etc

Data Mining Process



Big Data Technology Vendors



Big Data: The Moving Parts

Increasing
Age & Maturity

Hadoop

Vertica

MapReduce

Esper

kdb

Greenplum

ETL

ECL

Netezza

Teradata

Fast Data

Hive

SciPy

Mahout

MATLAB

Revolution R

SPSS

AMPL

SAS

Big Analytics

unsupervised learning

social media analytics

sentiment analysis

predictive modeling

BPO

BI

network analysis

visualization

simulation

Deep Insight

- mass customization of services
- quicker response to market trends
- identifying real-time cost optimizations
- faster, more accurate decision making
- better and more holistic R&D
- autonomic supply chain management

Business Objectives

From <http://blogs.zdnet.com/Hinchcliffe>

the growth of data will be exponential for the foreseeable future

terabytes

petabytes

exabytes

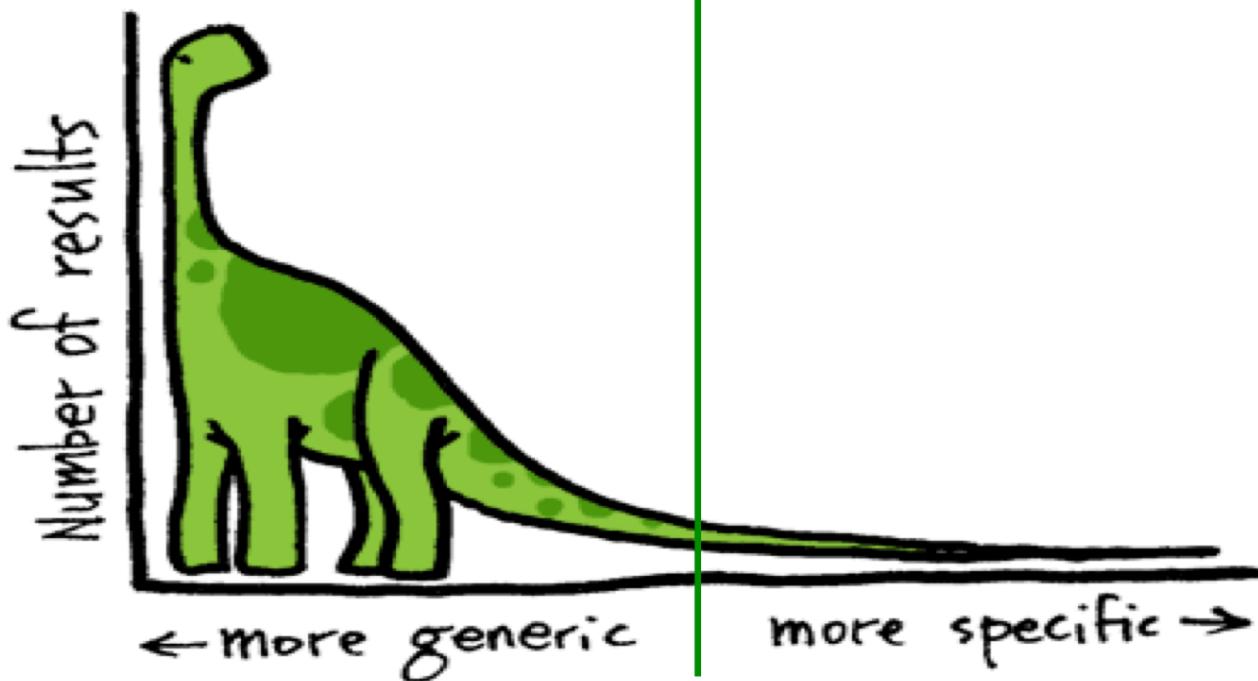
zettabytes

the amount of data stored by the average company today

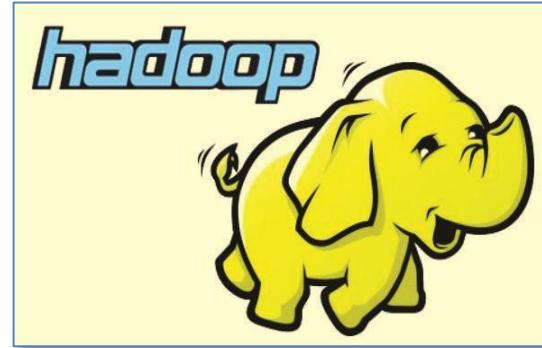
Big Data Science- A perspective

- Traditional DW
- Classical Stats
- Sampling

- Big Data
- Specific spikes
- Median is not the message



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

Scalable

Reliable

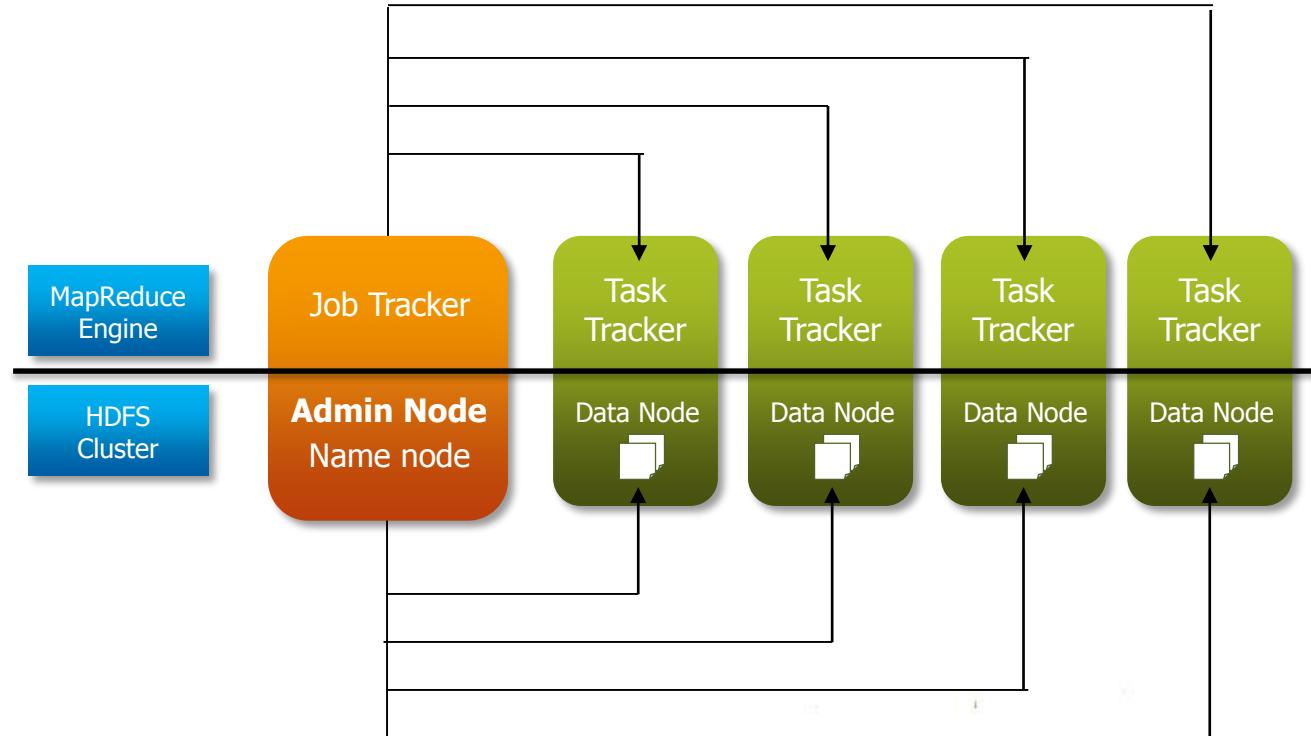
Economical

Flexible

Robust
Ecosystem

Real Time

Hadoop Core Components



Analytics with R



Analytics with R

The screenshot shows the homepage of The R Project for Statistical Computing. On the left, there's a sidebar with links for About R, What is R?, Contributors, Screenshots, What's new?, Download, Packages, CRAN, R Project Foundation, Members & Devs, Mailing Lists, Bug Tracking, Developer Page, Conferences, Search, Documentation, Manuals, FAQs, The R Journal, Wiki, Books, and Certification. The main content area features several data visualization examples: a PCA plot titled 'PCA 5 vars' showing variables like Family, Church, Education, and Agriculture; a clustering diagram titled 'Clustering 4 groups' showing four distinct clusters; a Factor Analysis plot titled 'Factor 1 [41%]' and 'Factor 3 [19%]'; and two density plots for 'Groups 1' and 'Groups 2'. At the bottom, a 'Getting Started:' section contains the following text:

• R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred CRAN mirror.

• If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

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

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

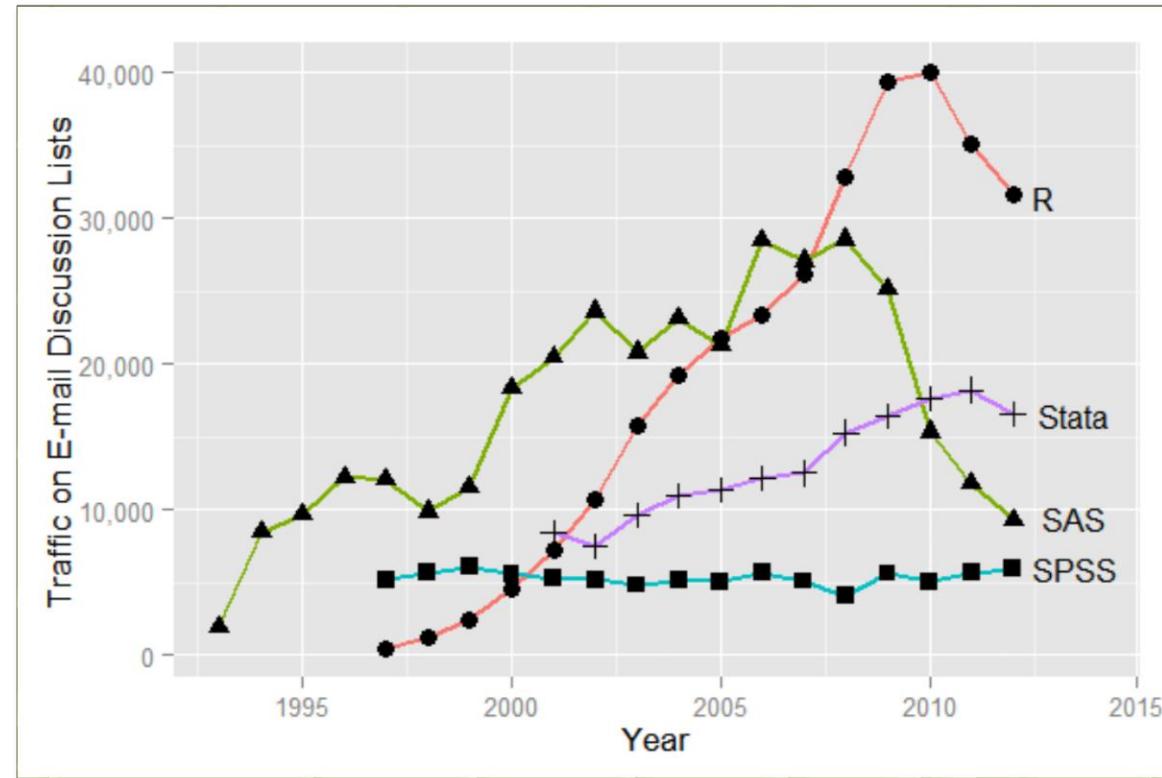
Customization needs **command line**

SAS GUI are more expensive

Comparing R

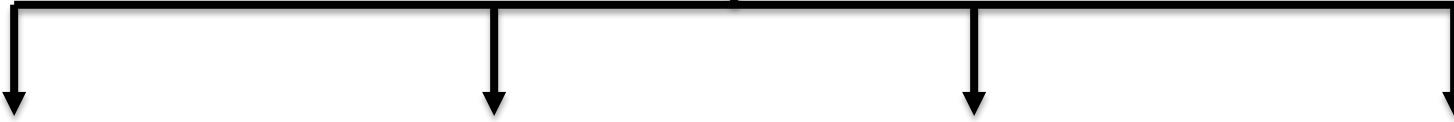
Comparing R and others

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



Introduction to R Programming language

www.r-project.org/about.html



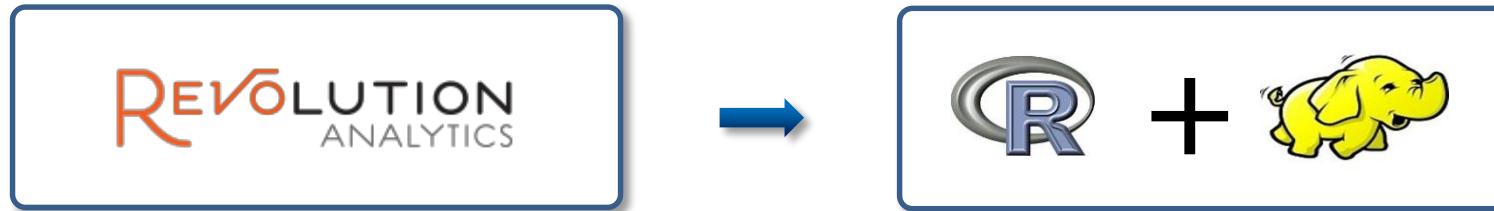
- ✓ History
- ✓ Evolution
- ✓ Current state

- ✓ Open Source
- ✓ Free
- ✓ Widely Recognized

- ✓ Official Website
- ✓ R Core
- ✓ Creators

- ✓ R Journal

R and Hadoop Integration

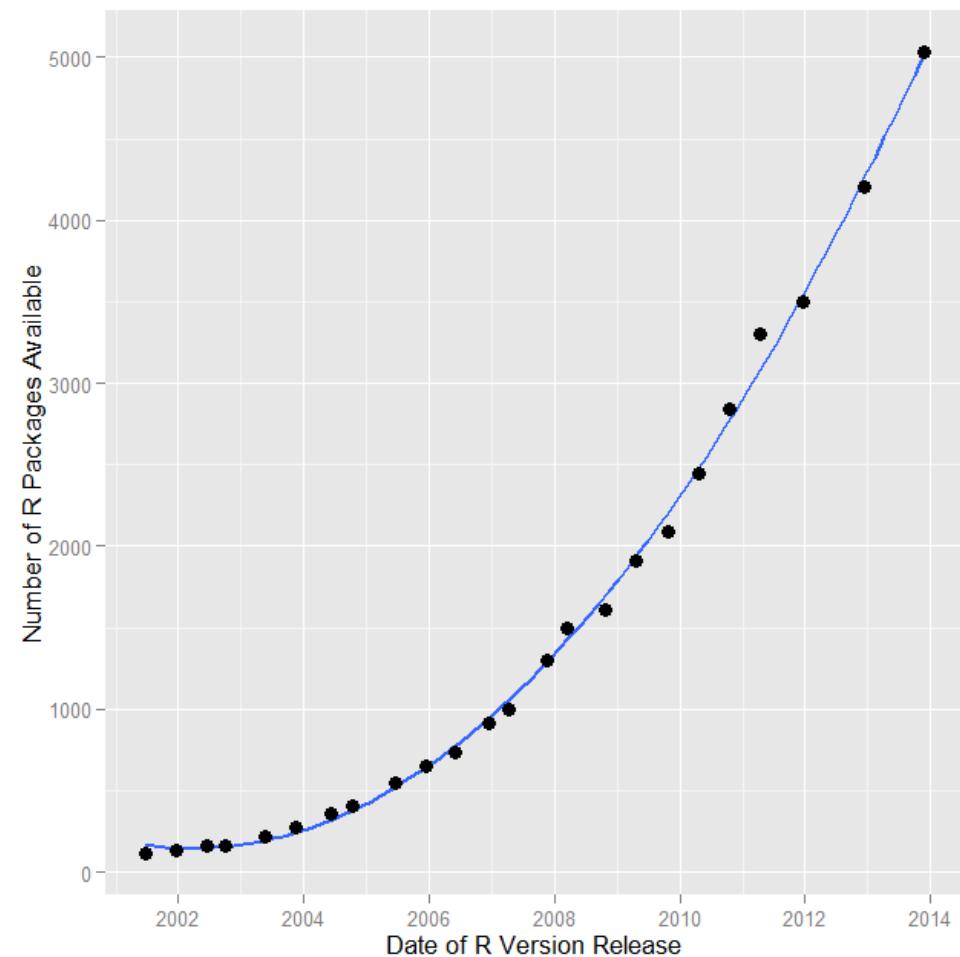


- R and Hadoop are a natural match in Big Data Analytics and visualization.
- One of the most well-known R packages to support Hadoop functionalities is : **RHadoop**
- Rhadoop was developed by [RevolutionAnalytics](#).
- RHadoop is a collection of three R packages: `rnr`, `rhdfs` and `rbase`
- rnr package provides Hadoop MapReduce functionality in R, rhdfs provides HDFS file management in R and rbase 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?



Open-source stat package with visualization

Source: www.r-project.org

Vibrant community support.

One-line calculations galore!

Steep learning curve but worth it!

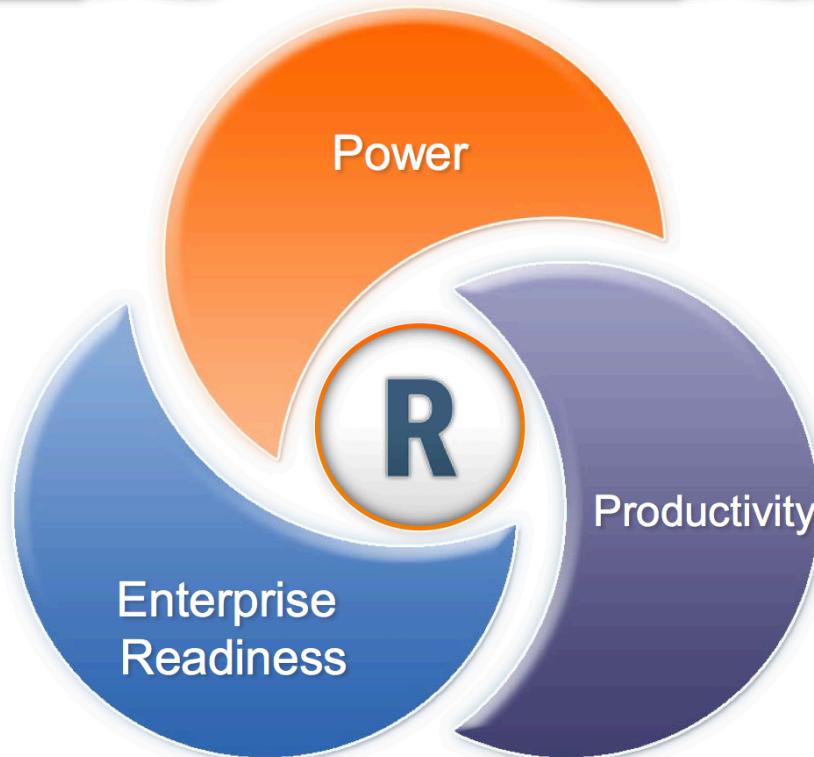
Insight into statistical properties and trends...

or for machine learning purposes...

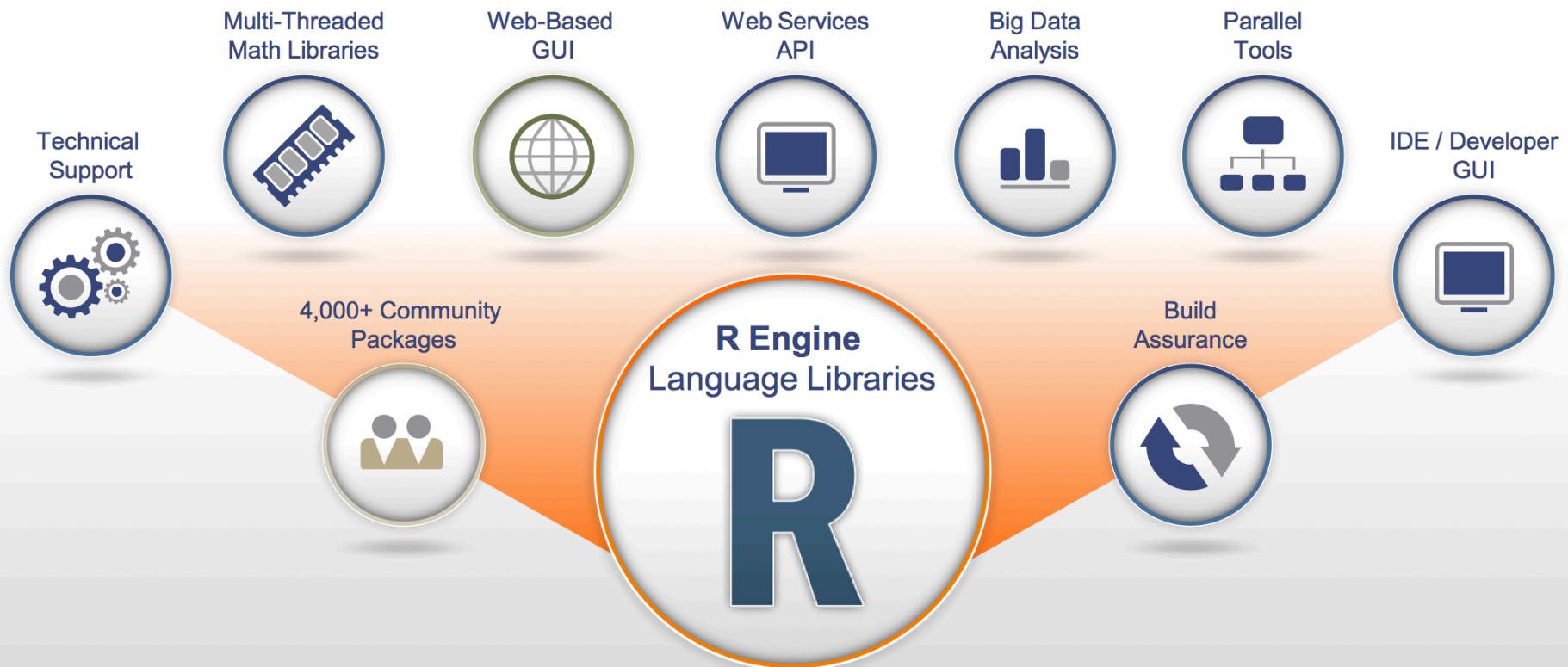
or Big Data to be understood well.

R Open Source Analytics

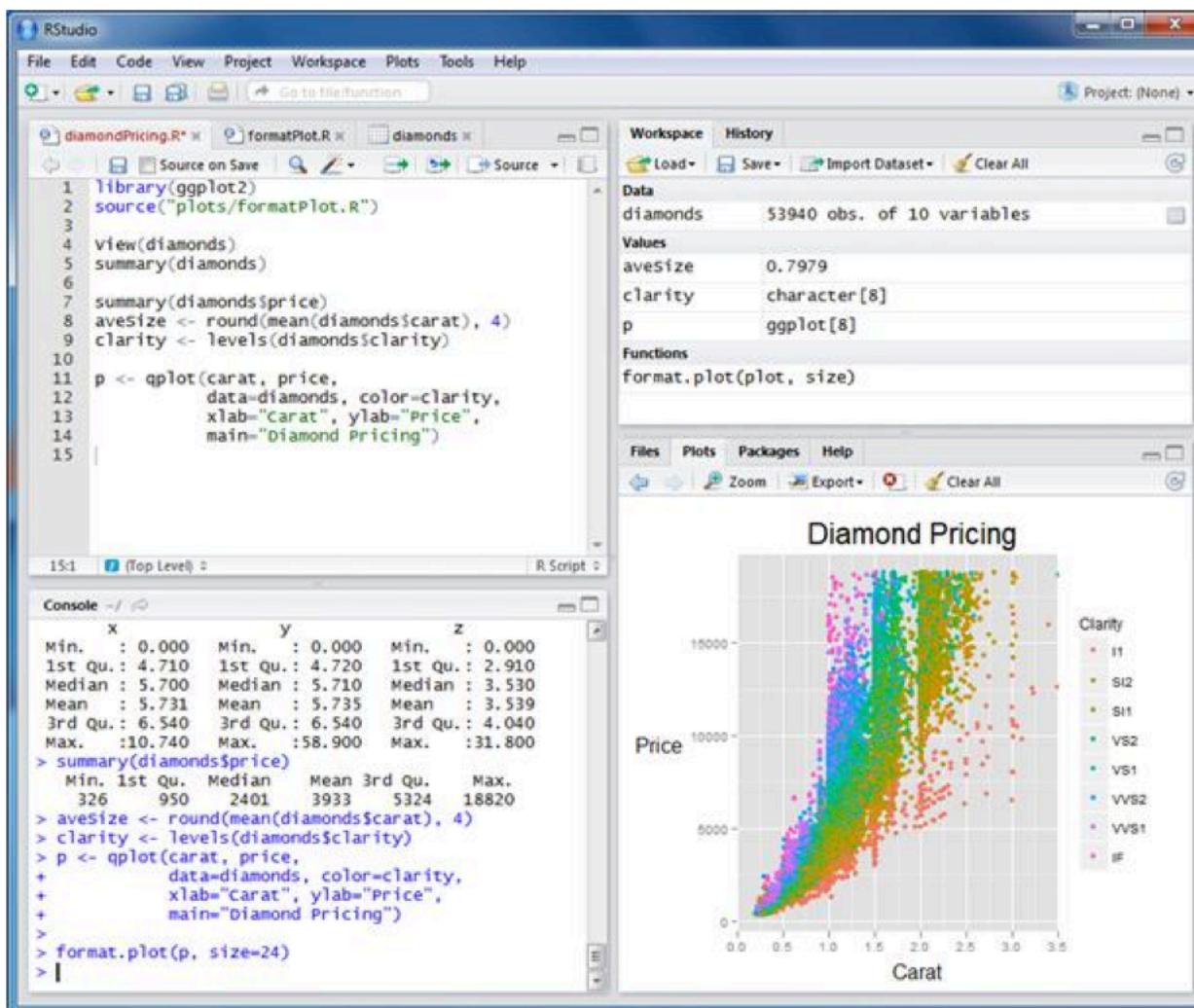
- Most advanced statistical analysis software available
- Half the cost of commercial alternatives
- 2M+ Users
- 3,000+ Applications



R Universe



R-Studio



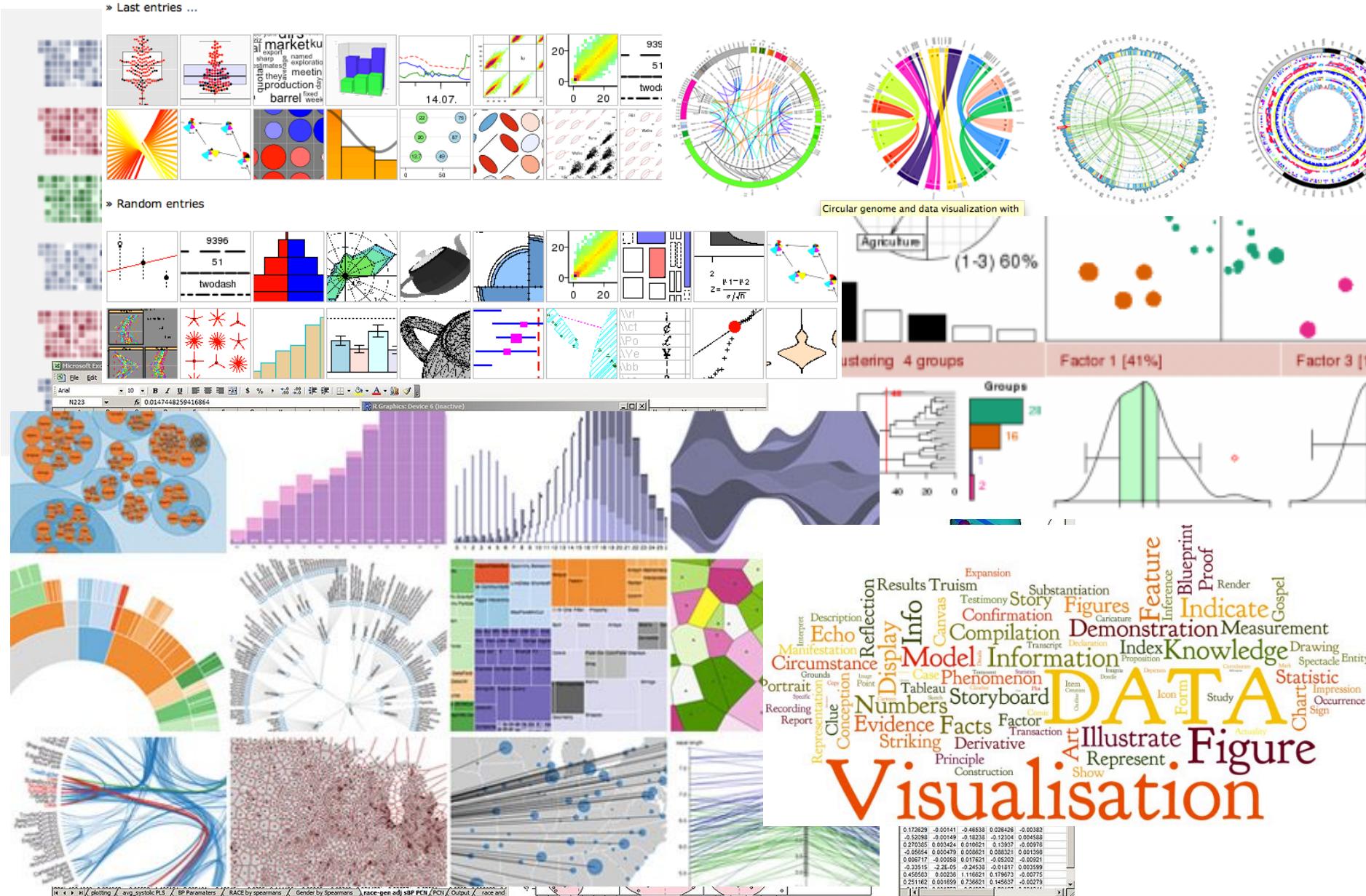
Source: www.rstudio.com/ide/

CASE STUDIES AND DEMO USING R



Source: www.r-project.org

Examples of R Visualizations



Machine Learning



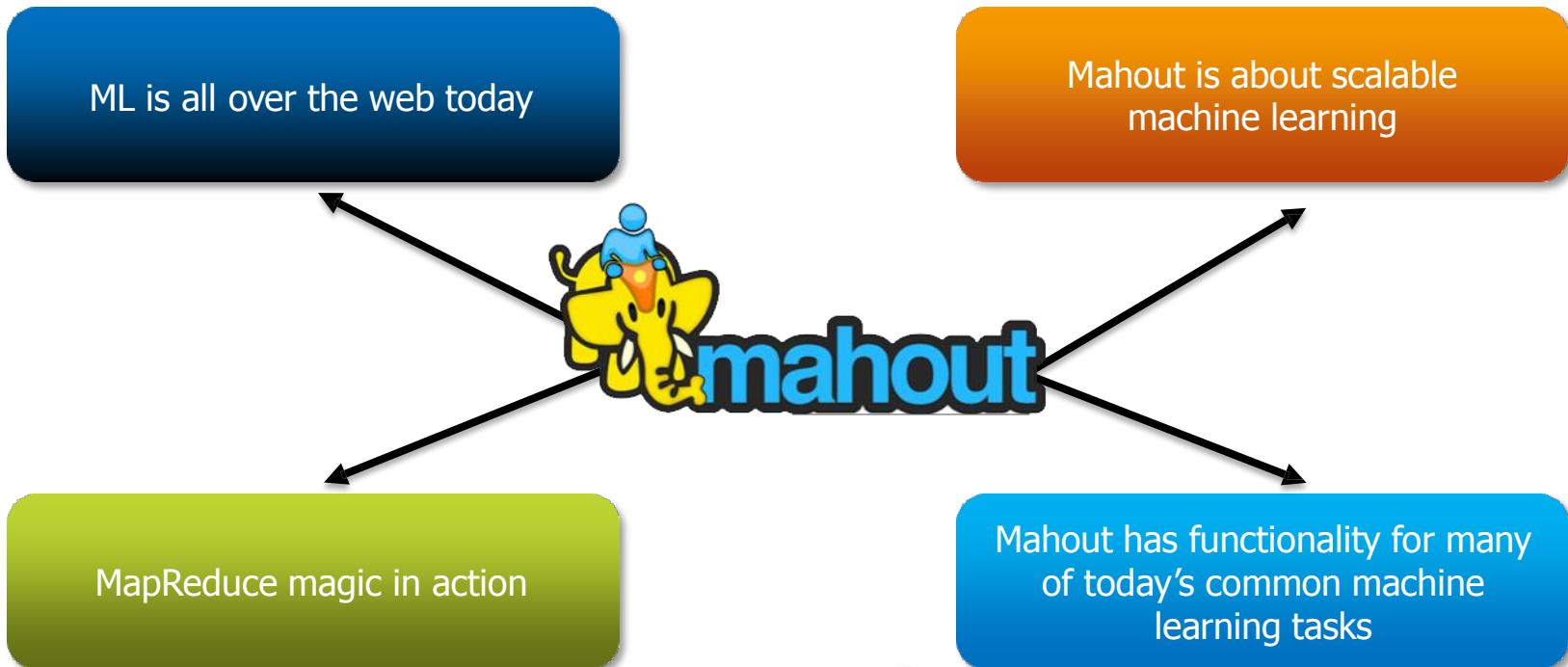
Machine Learning: Mahout

- ✓ **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



Machine Learning with Mahout



Write intelligent applications using Apache Mahout

Hadoop and
MapReduce magic in
action

COMPANIES YOU MAY WANT TO FOLLOW

The image shows a grid of nine company logos. The first row contains HCL (blue text), Capgemini (blue square with white cloud), and HEWLETT (yellow cube). The second row contains CISCO (blue square with white bars) and DELL (blue circle with white letters). The third row contains TERADATA (orange text) and Apple (white silhouette).

LinkedIn Recommendations

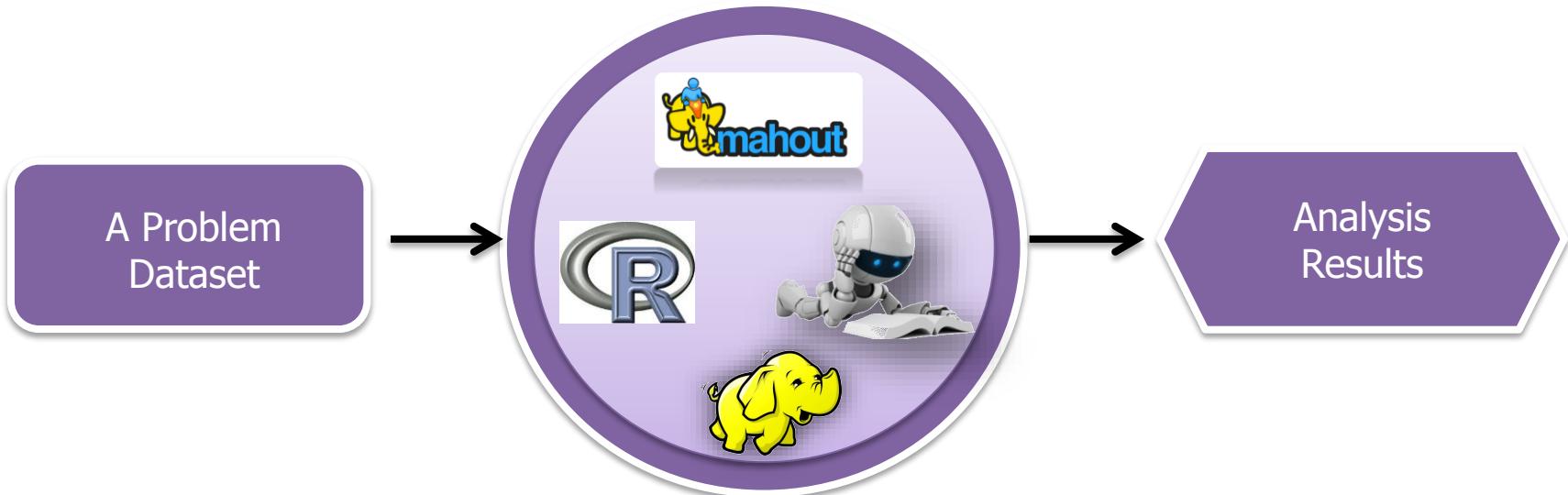
JOBS YOU MAY BE INTERESTED IN

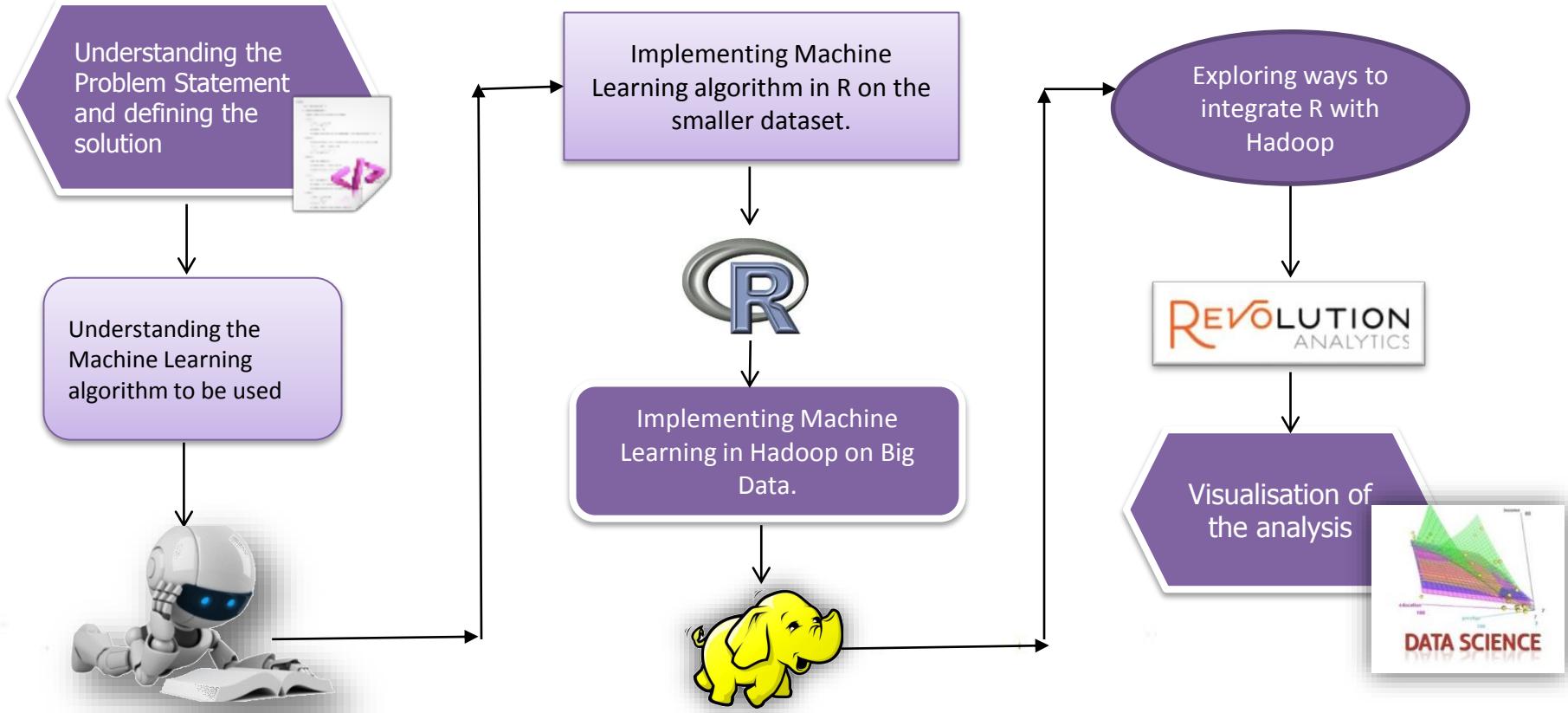
- HCL** Manager Global Infra Pre Sales
HCL Technologies - Noida , ...
- Pivotal** Sr Account Manager, Pivotal
Pivotal Inc. - India - Mumbai
- TERADATA** Solution Sales Specialist
Teradata - IN-Karnataka-Ban...

[Feedback | See more »](#)

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

Datasets





Dataset 1

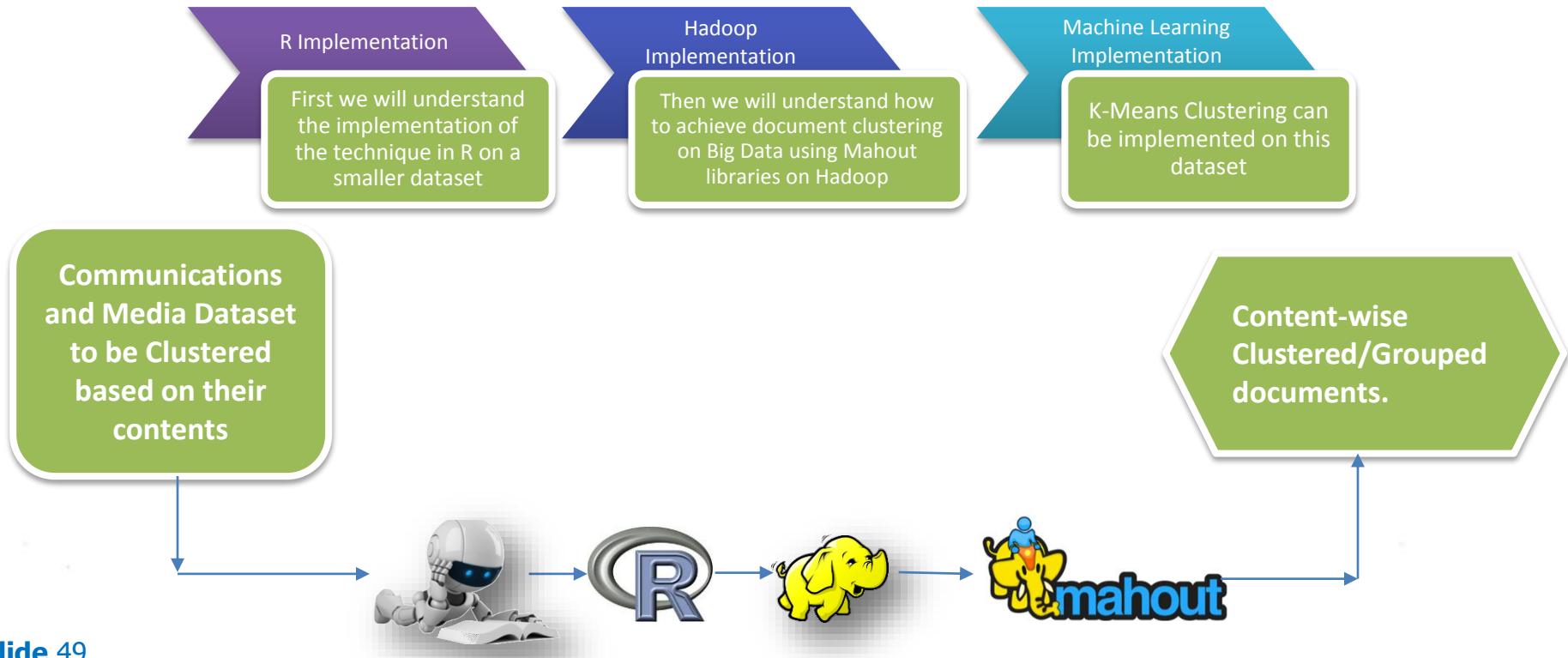
Domain of the Dataset:

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.



Dataset 2

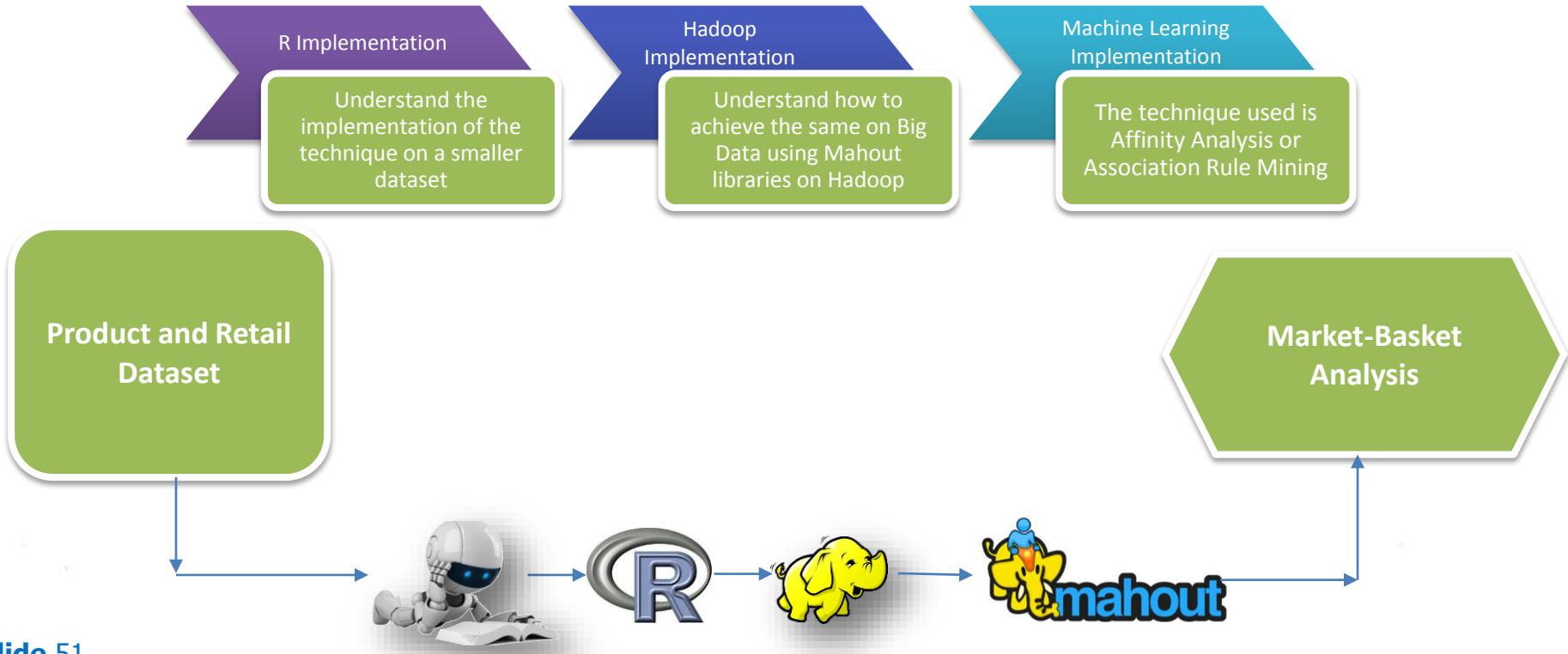
Domain of the Dataset:

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.



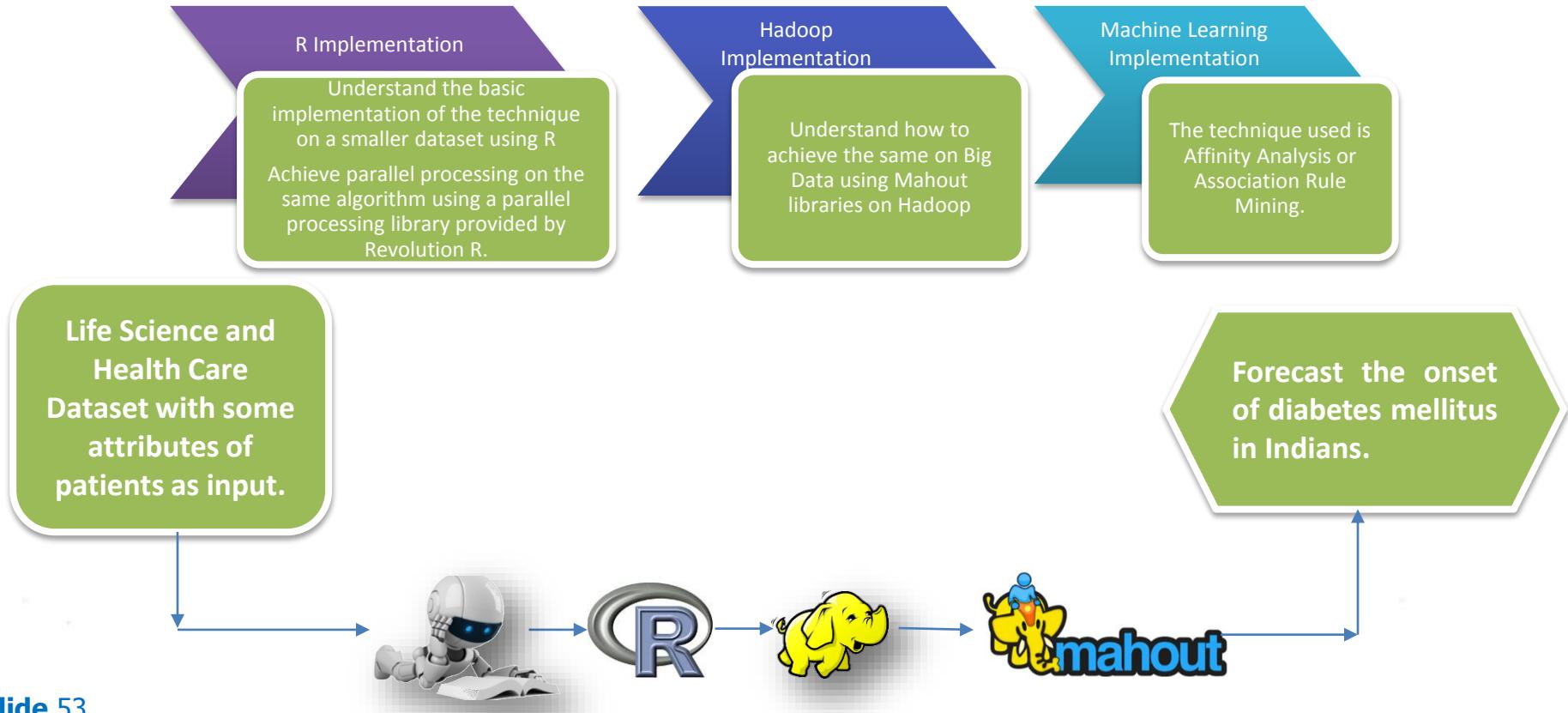
Dataset 3

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.



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, ...)

