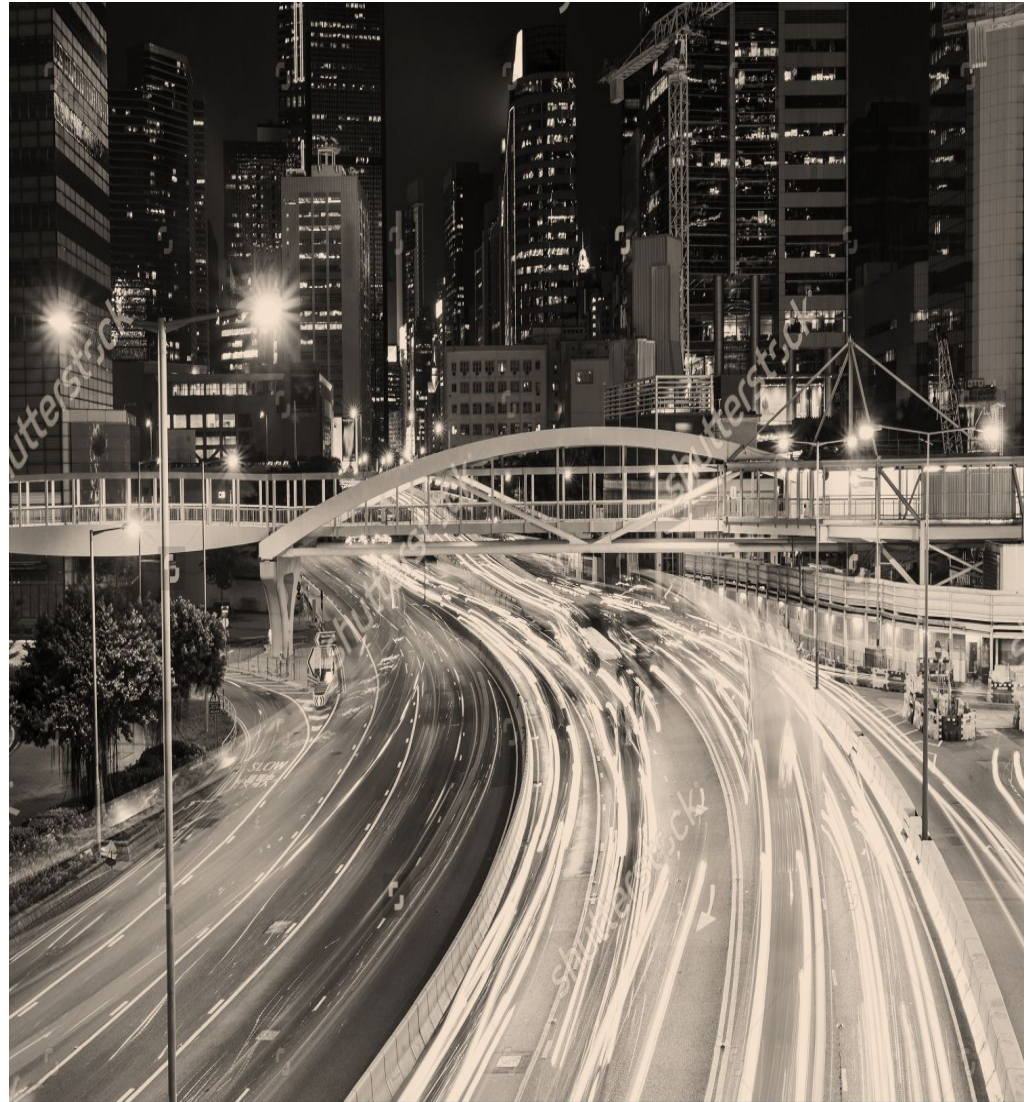


Moving Beyond Business Intelligence



shutterstock

IMAGE ID: 209821609
www.shutterstock.com

Evolution of Analytics



Looks forward at the future

Streaming Analytics

Deep Learning , Artificial Intelligence

Predictive Analytics

Building Models , Test them

Looks at what happened in the past

Business Intelligence

Visualization,
OLAP,
Dashboard

Intuition , domain
knowledge



Insights

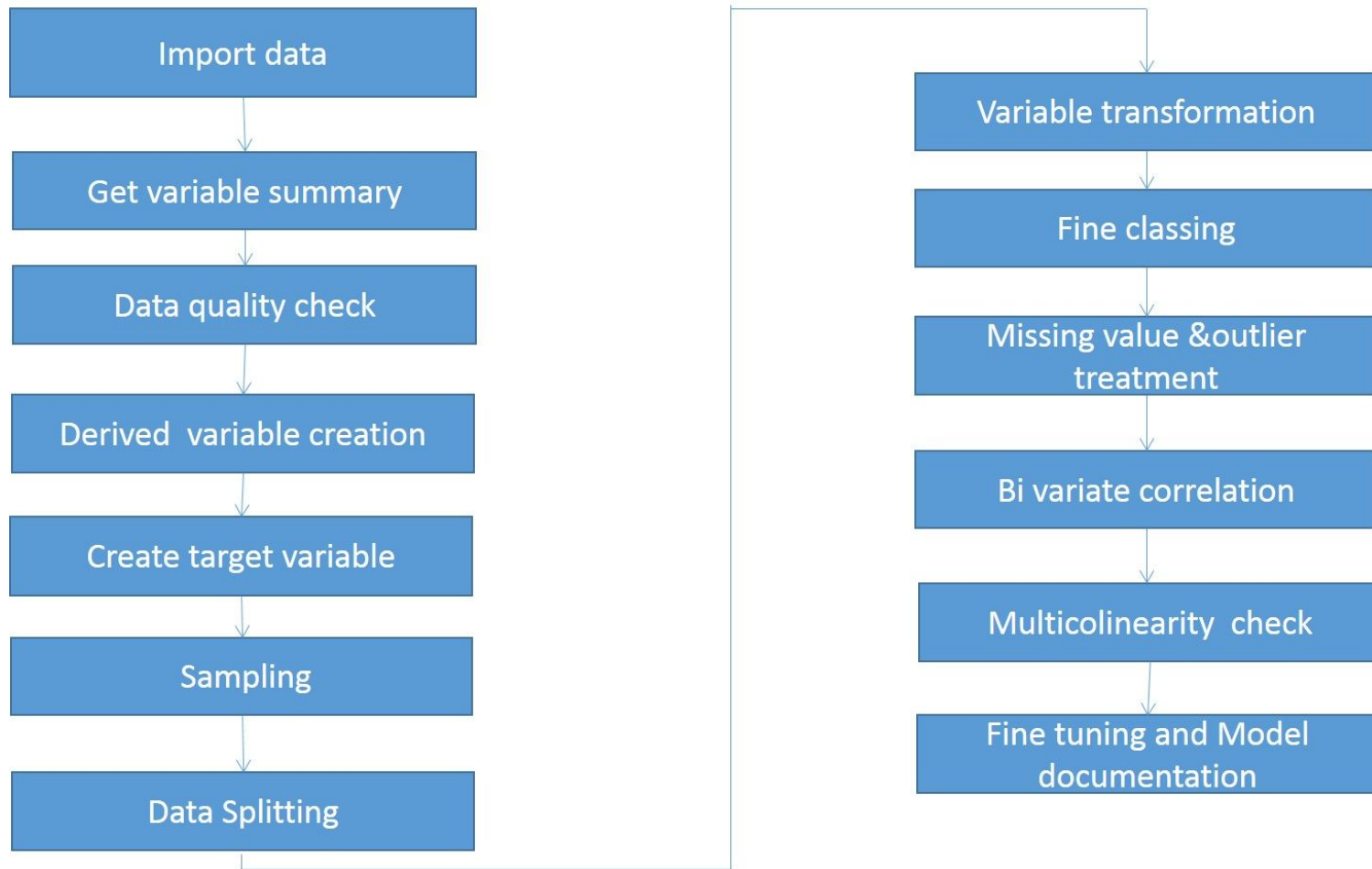
Complexity

Predictive Analytics

Predictive Analytics an area of data mining that deals with extracting information from the data and use it to predict trends and user patterns.

----Wikipedia

Not an easier workflow in real time



Traditional Tools

- Open Source - R, Python
- Commercial tools – SAS, SPSS, SATA , Excel, MinTab

Techniques

- Regression analysis
 - Linear, Logistic regression
- Statistical tools , ANOVA

Applications

- Weather Prediction
- Election polls (Exit-Polls)
- Healthcare treatments
- Advertising Companies
- On-line portals, Ecommerce based applications
- Financial Sector

Limitations:- Data should be collected intentionally,
Very Expensive!!!

Big Data

High-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

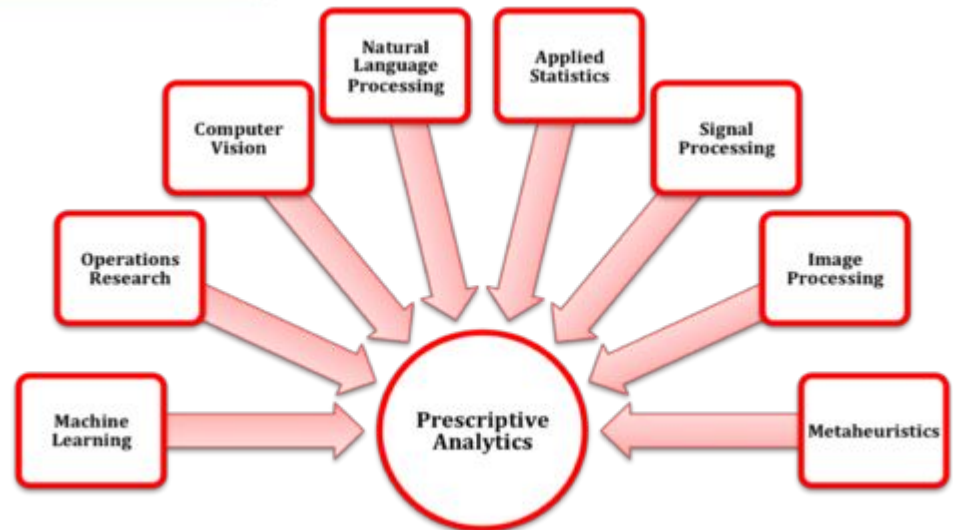
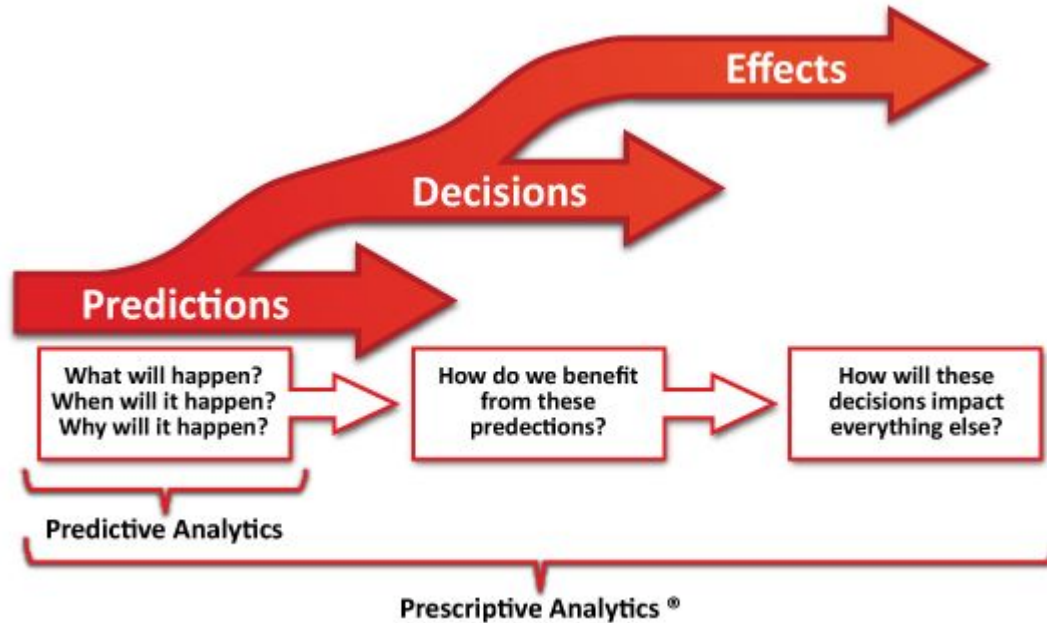
Doug Laney, Gartner

Big Data Content

- Social Networking sites
- Logs
 - Web server logs
 - Audit logs
- Internet Content
- IoT (Internet of Things)
- Open data movement (academia, government)



Prescriptive Analytics



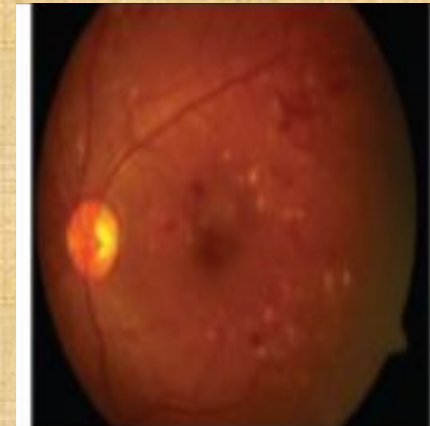
Use cases – Prescriptive Analytics

Preventing Derailment



- Use Information from wheel sensors
- Analyze the data
- Take prescriptive actions:
 - stop the train & intimate the maintenance crew

Diagnosis of Diseases



- Extract the features of interest
- Find the severity of the disease
- Suggest immediate prescriptive actions

Big Data Tools

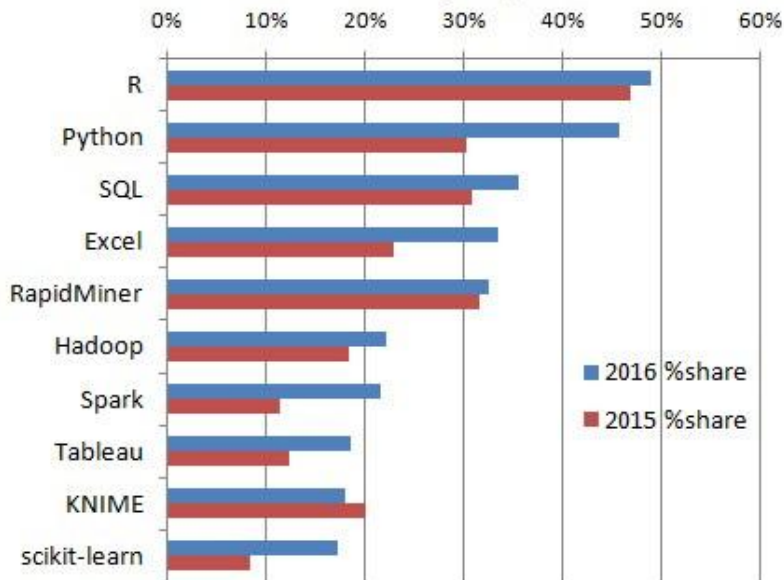
- Open Source Tools

- Spark/MLlib
- Hadoop + Python Scripts
- R + Traditional Tools (If the data can fit into single processor)

- Commercial Tools

- HP Vertica distributedR
- RapidMiner (Radoop on Hadoop analytics)
- SAS Enterprise Miner

KDnuggets Analytics/Data Science 2016 Software Poll, top 10 tools

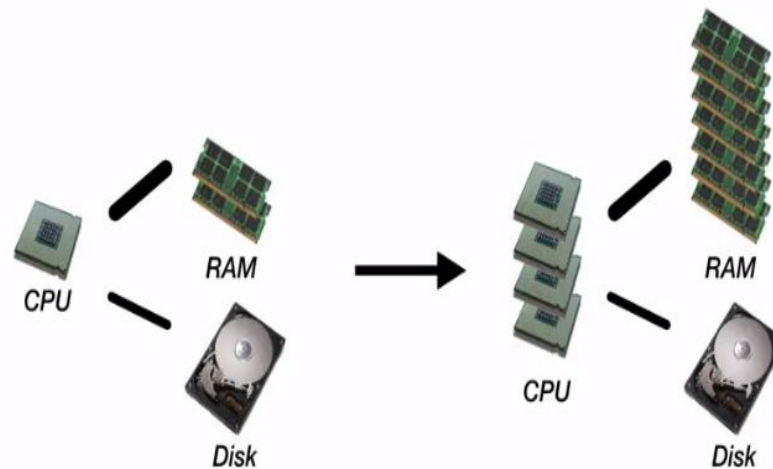


Tool	2016 %Share	2015 %share	% change
Hadoop	22.1%	18.4%	+20.5%
Spark	21.6%	11.3%	+91%
Hive	12.4%	10.2%	+21.3%
MLlib	11.6%	3.3%	+253%
SQL on Hadoop tools	7.3%	7.2%	+1.6%
H2O	6.7%	2.0%	+234%
HBase	5.5%	4.6%	+18.6%
Apache Pig	4.6%	5.4%	-16.1%
Apache Mahout	2.6%	2.8%	-7.2%
Dato	2.4%	0.5%	+338%
Datameer	0.4%	0.9%	-52.3%
Other Hadoop/HDFS-based tools	4.9%	4.5%	+7.5%

Distributed Computing

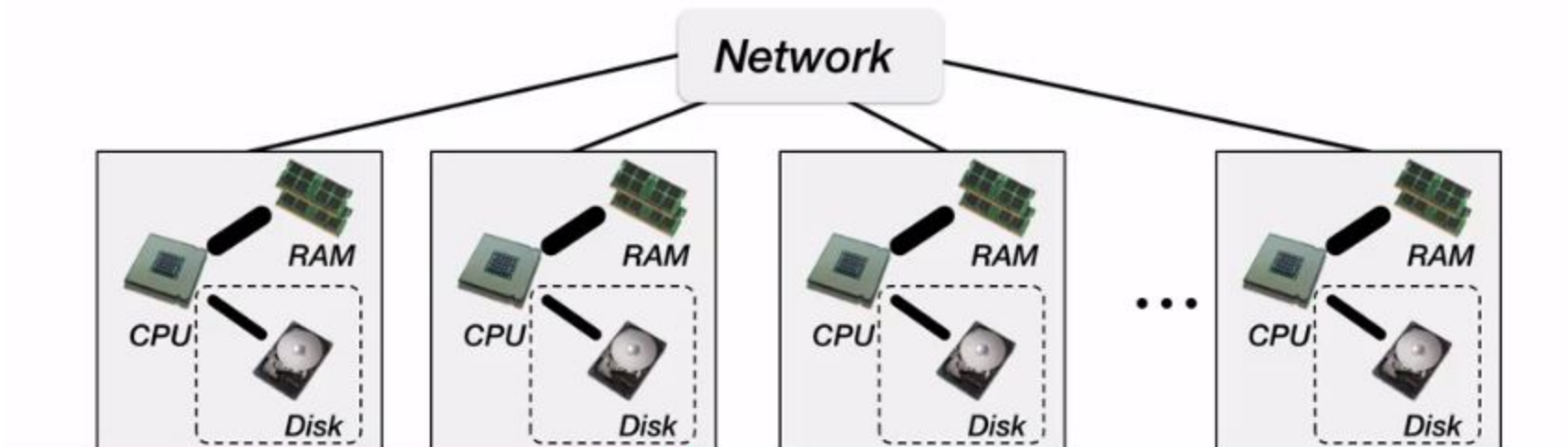
- Why can't Traditional Tools (Matlab, R, Excel) cannot be used for processing large datasets
 - They typically run on single machine.
 - Need more hardware to store / process data

- **Scale – up** the machine (large machine)
 - Good Idea ! Actually it works faster
 - But need Specialized hardware (expensive)
 - Scaling can be done to a certain extent

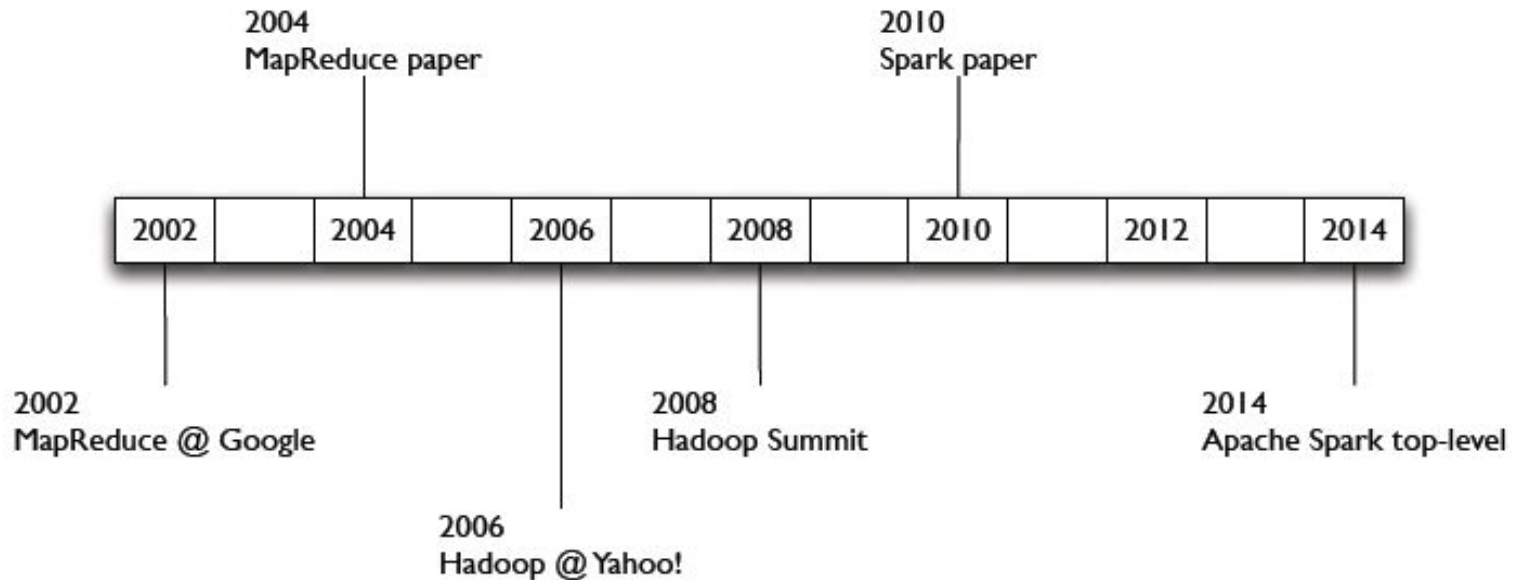


- **Scale out**

- Many small machine connected them over network in distributed setting
 - Better alternative , as nodes can easily be added
 - Commodity hardware
 - Network Communication , Software Complexity



Cluster Computing Platforms



Hadoop vs. Spark

- Two main challenges with Hadoop
 - Configuring Cluster
 - Complex Programming model (MapReduce)

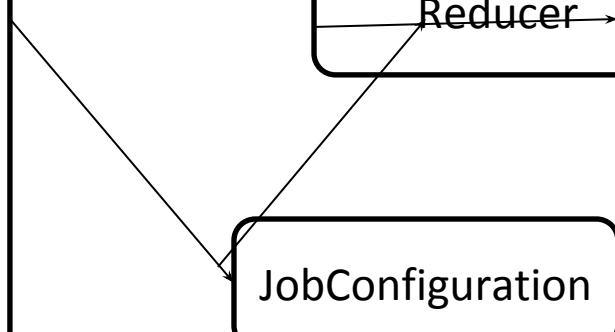
```
void map (String doc_id, String text)
for each word w in segment(text)
emit(w, "1");
```

```
void reduce (String word, Iterator
group):
int count = 0;
for each pc in group:
count += Int(pc);
emit(word, String(count));
```

Mapper

Reducer

JobConfiguration





- Open source cluster computing engine
- Why spark for large scale machine learning?
 - Fast iterative computation
 - Communication primitive
 - Provides API for scala, python and java
 - Interactive shell
 - Many high level libraries are available for building machine learning pipelines

Big Data Applications

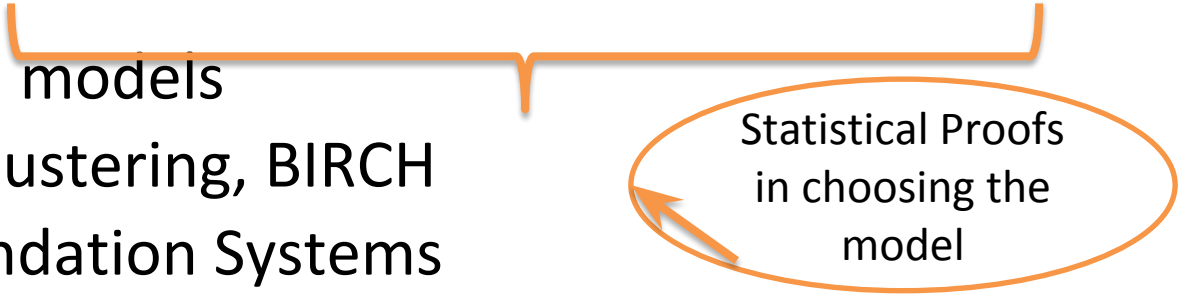
- Recommendation systems
- Spam Filtering
- Speech Recognition
- Face Recognition
- Link Prediction
- Protein Structure prediction

Machine learning

Constructing and studying methods that learn from
and make predictions on data

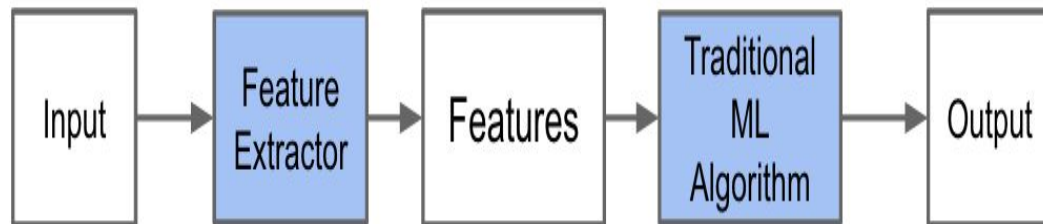
- Methods

- Random Forest Trees, Naive Bayes, SVM, Ensemble models
- Regression models
- K-means clustering, BIRCH
- Recommendation Systems

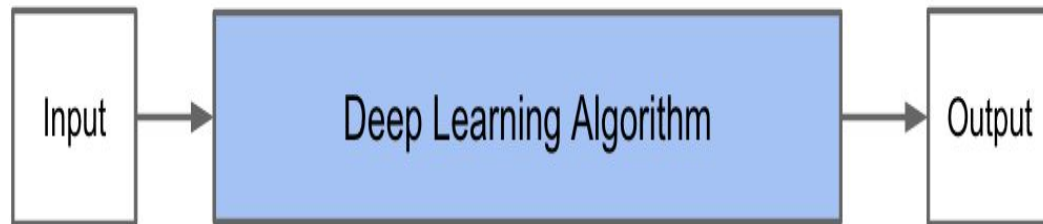


Statistical Proofs
in choosing the
model


Deep Learning



Traditional Machine Learning Flow



Deep Learning Flow

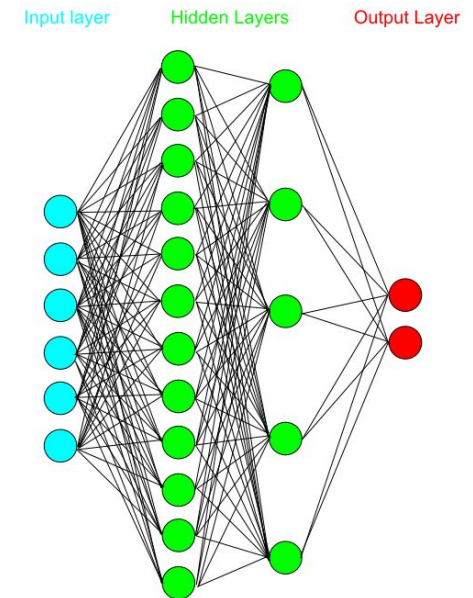
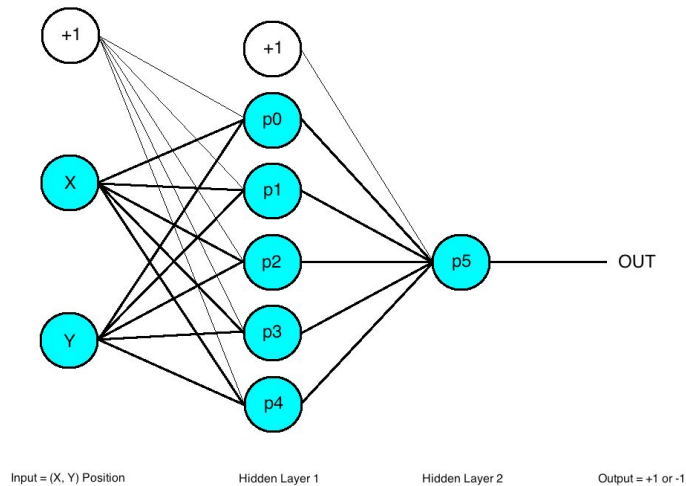
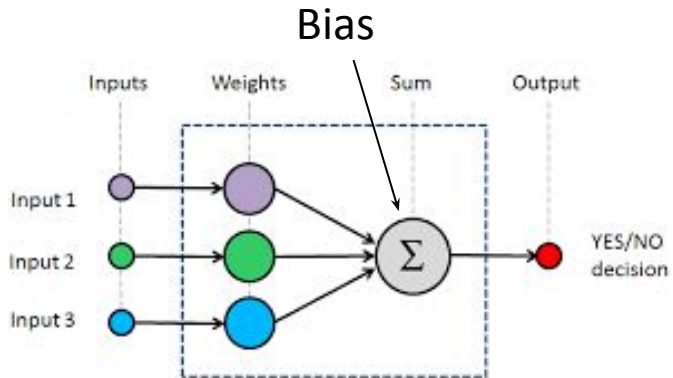
The background of the slide features a dark blue field with various financial charts. In the upper right, there is a candlestick chart with green and red bars. Below it, a yellow line chart trends upwards. In the lower right, a blue line chart with green and red circular markers is visible. On the left side, there is a vertical list of financial data points in white text.

**By 2020, 20% of
companies will dedicate
workers to monitor and
guide **neural networks**.**

gartner.com/SmarterWithGartner

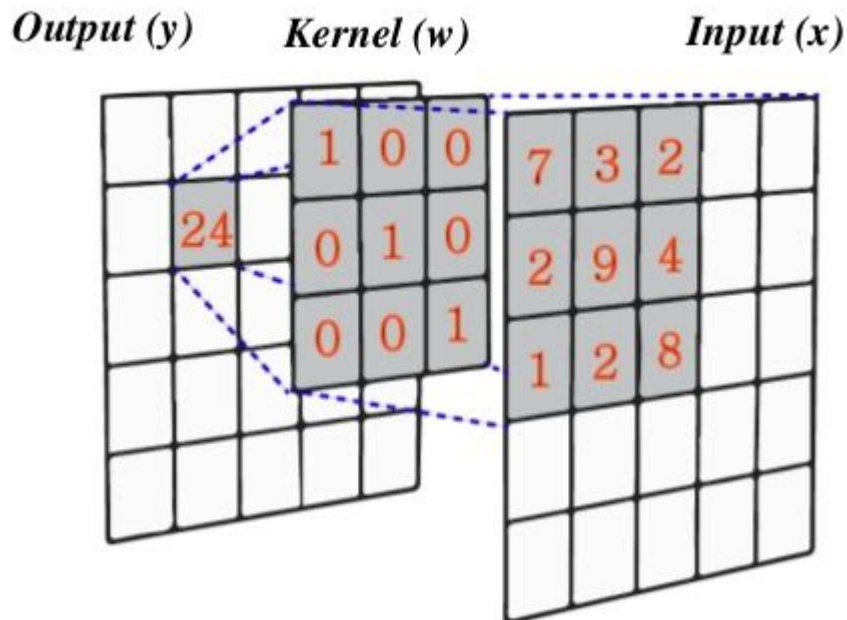
Gartner

Neural Networks

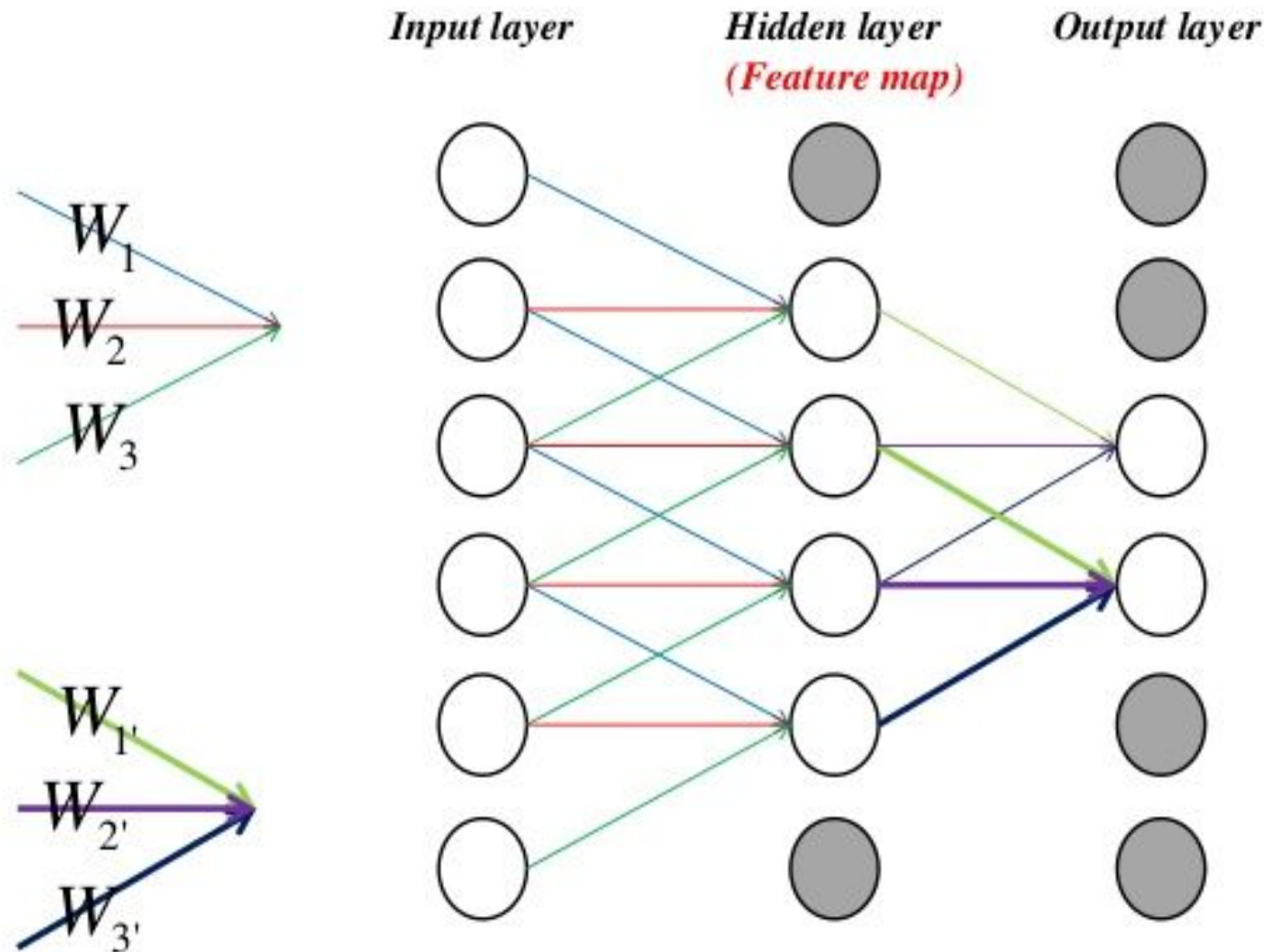


Convolutional Neural Networks

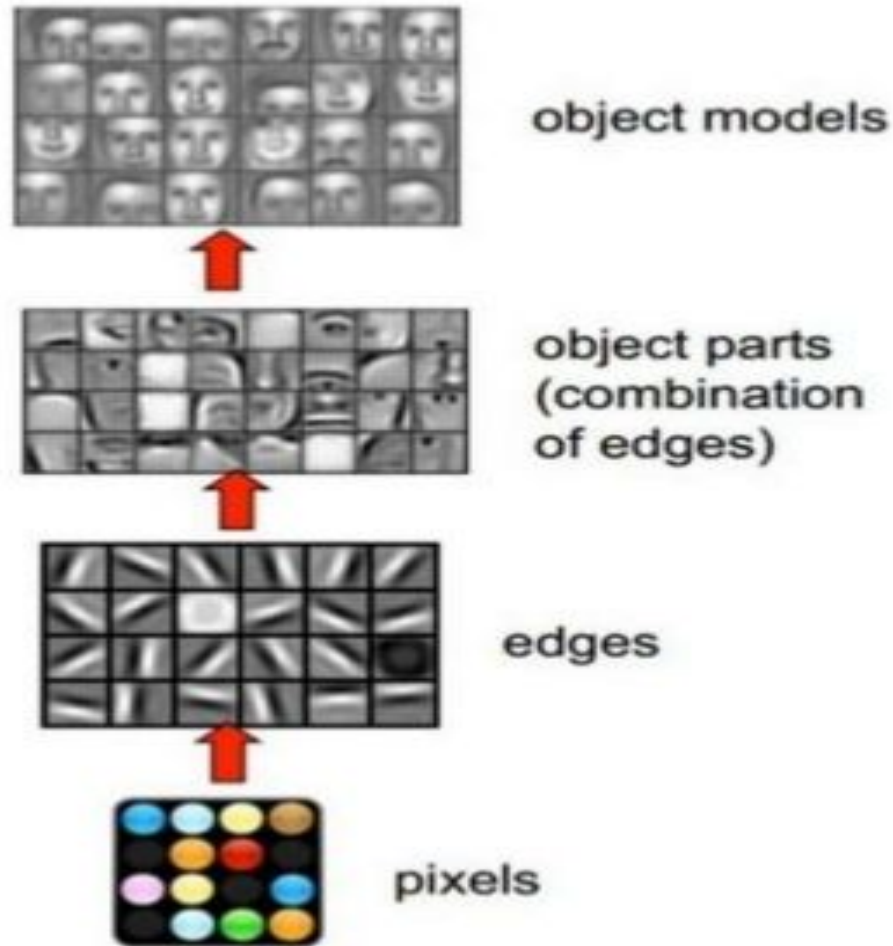
- Weights are shared by neurons
- Makes use of local neighbourhood in the images



Shared Weight representation



CNN for Face Recognition



Visualization

"The key to effective analysis of big data is the integration of visualization into analytics tools so that all kinds of users can interpret big data from a wide range of sources—clickstreams, social media, log files, videos and more." *IBM*

Harvard
Business
Review

The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions



Study of **visual representation of data** meaning information that has been abstracted in some schematic form including attributes or variables for the units of information

Wikipedia

Challenges

- Demand for easy-to-use / self service tools
- Managing data quality
- Embed Analytic and Visualization tools onto mobile platforms

Previous Works

- Touch Dynamics
- Lead Prioritization using Logistic Regression
- Twitter stream analysis using storm
- Web log analysis using spark
- Parallel Matrix Factorization

Current Works

- Disease Diagnosis using Deep Neural networks
- Face Recognition using Transfer Learning

Hope the session was useful

Thank You!