Advanced Analytics:
Machine Learning with R and Python

Dr. Ash Pahwa

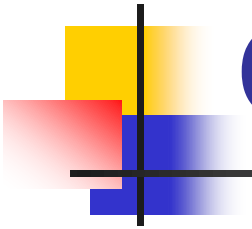
California Institute of Technology
Center for Technology and Management Education
1200 East California Blvd., Mail Code 121-79
Pasadena, California 91125
Phone: 626.395.4042

Advanced Analytics: Machine Learning with R and Python

Lesson 1.0: Machine Learning and Predictive Analytics

Lesson 1.2

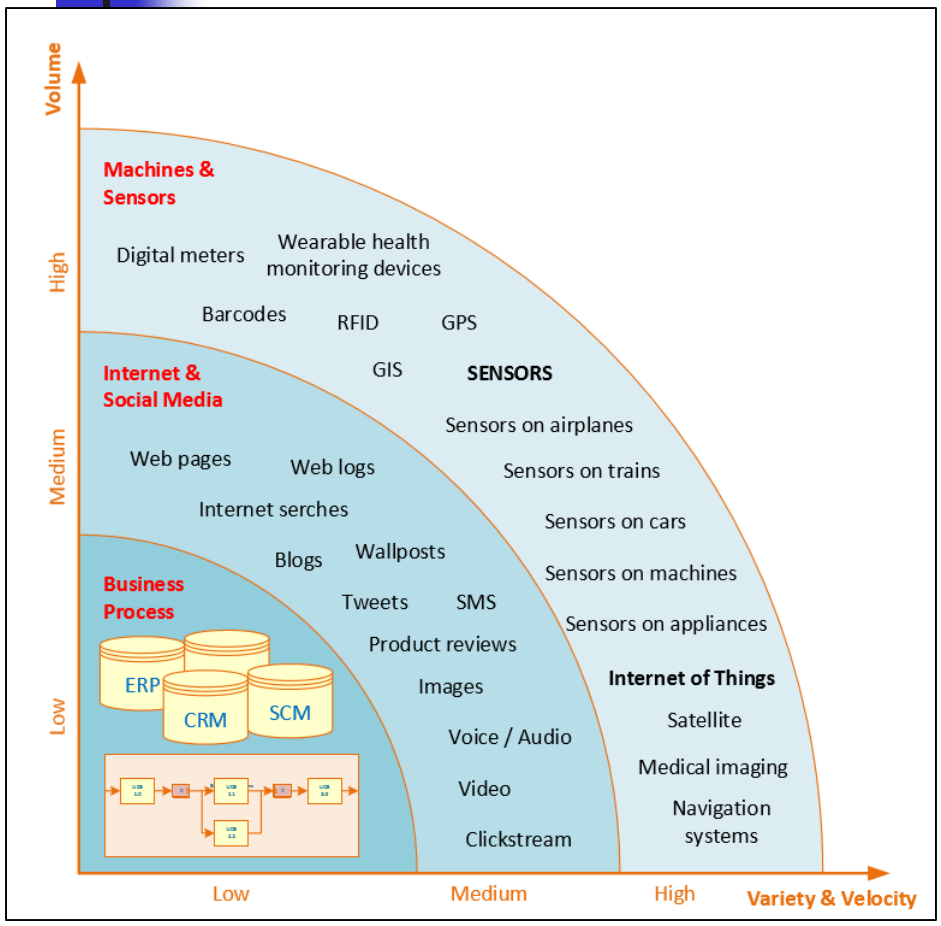
What is Machine Learning & Predictive Analytics?



Outline

- Big Data and Analytics
- Data Science Applications
- Trends in Technology

Big Data
Where does it come from?



Name	Symbol	Value
Kilo Byte	kB	10^3
Mega Byte	MB	10^6
Giga Byte	GB	10^9
Tera Byte	TB	10^{12}
Peta Byte	PB	10^{15}
Exa Byte	EB	10^{18}
Zetta Byte	ZB	10^{21}
Yotta Byte	YB	10^{24}
Bronto Byte	BB	10^{27}
Gego Byte	GeB	10^{30}

Other Names of Data Science

- Data Science
  - Machine Learning
    - Machine learns from the training data
    - Classify an entity by analyzing new entity's data
  - Data Mining
    - Data has many patterns
    - Data mining allows us to see those patterns
  - Predictive Analytics
    - Predicts the result given new data

Data Science

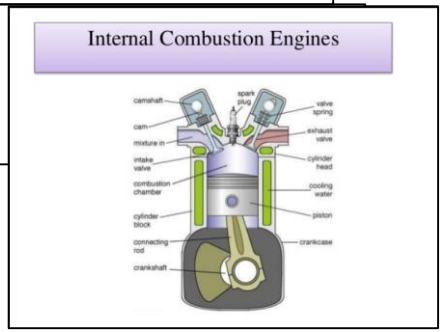
Motor Cars

Trucks

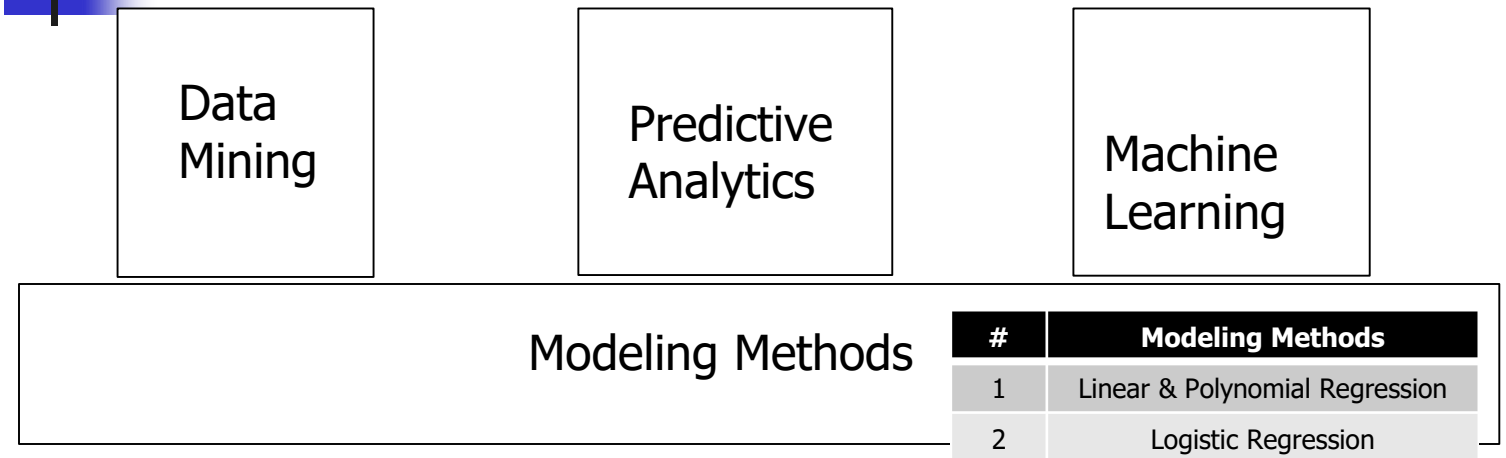
Bus + Vans

Earth Moving Equipment Caterpillar

Internal Combustion Engine

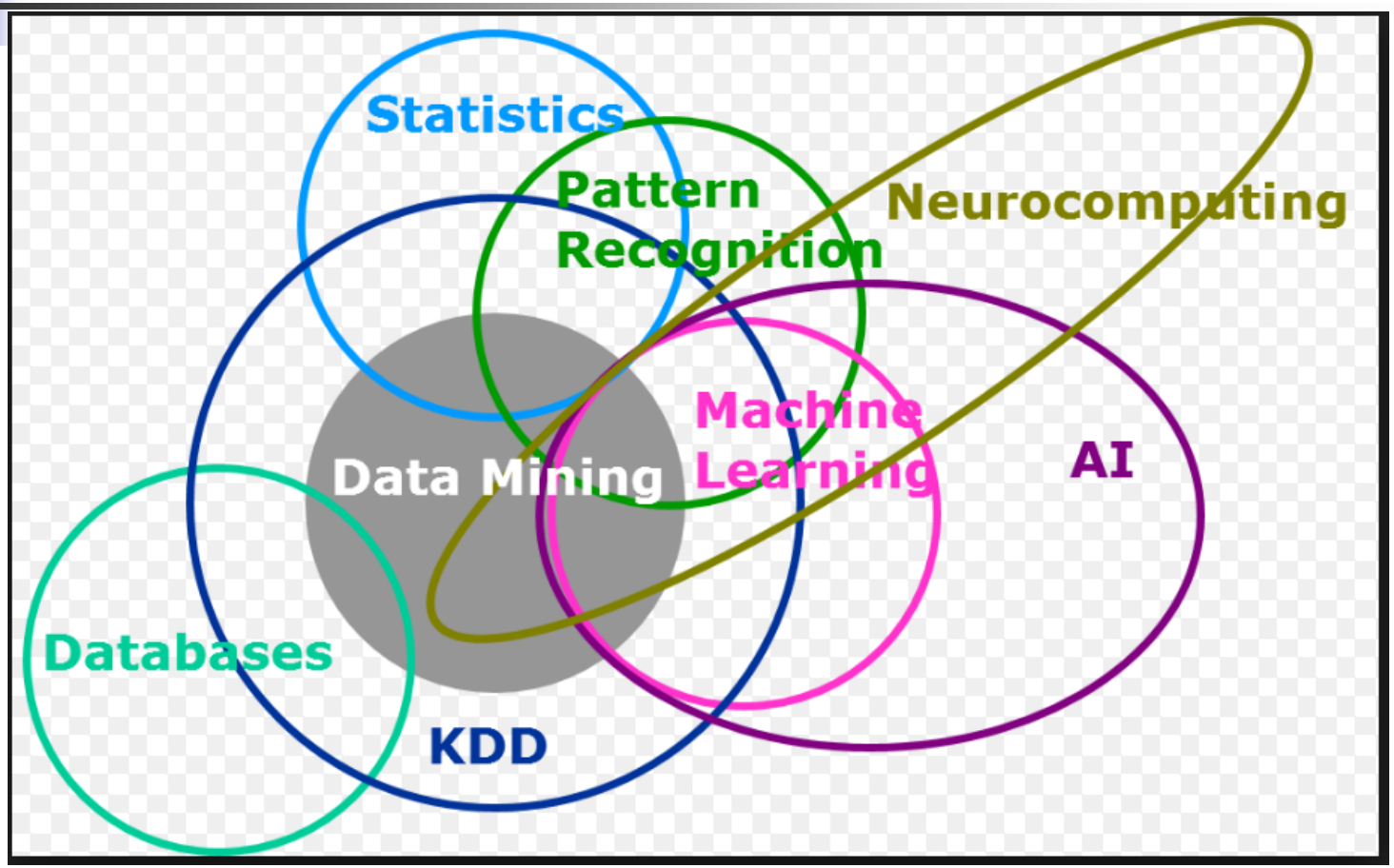


Data Science

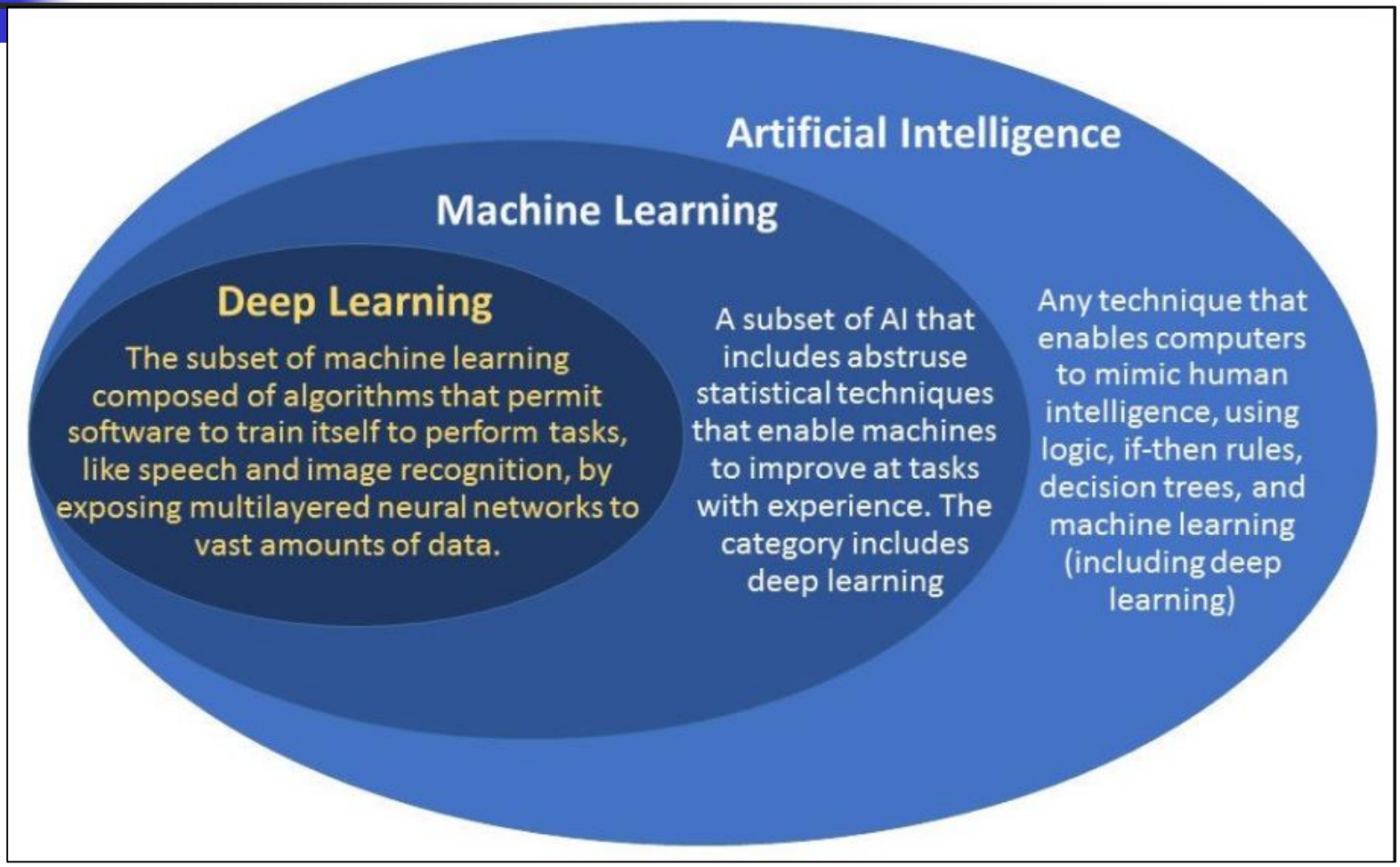


#	Modeling Methods
1	Linear & Polynomial Regression
2	Logistic Regression
3	Discriminant Analysis
4	K Nearest Neighbor
5	Decision and Regression Trees
6	Naïve Bayes
7	Neural Networks
8	Clustering
9	Principal Component Analysis
10	Support Vector Machines
11	ARIMA : Time Series

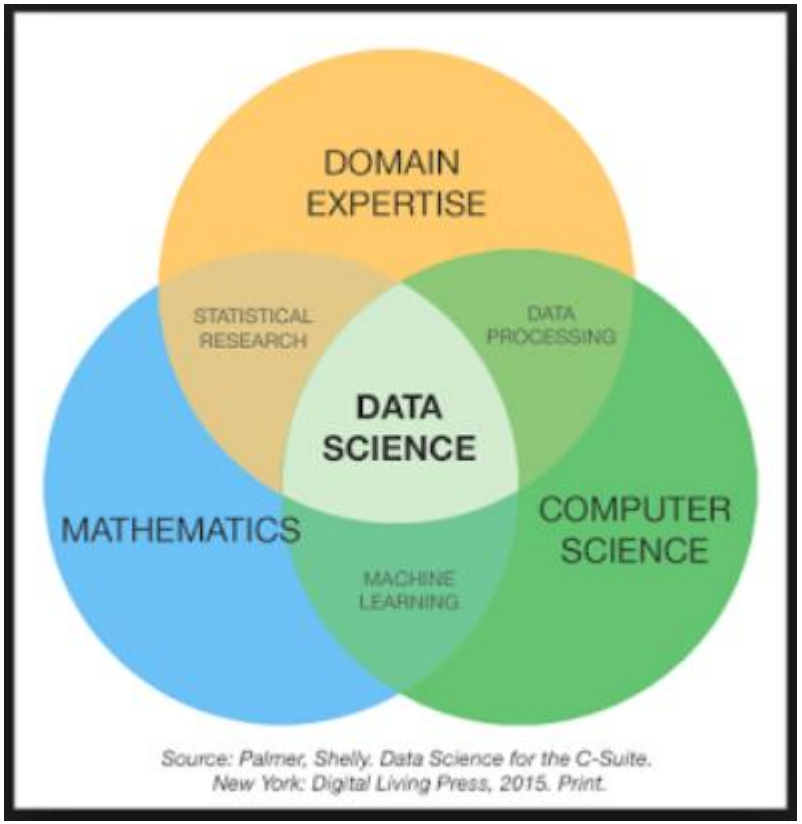
Data Science



Artificial Intelligence



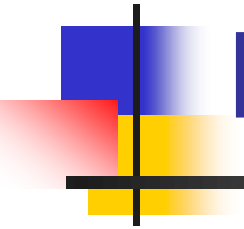
Data Science



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

New Technologies that Enable Data Science

- Hardware
  - High speed internet & wireless (mobile technology)
  - Access to web from smart phones – iPhone, Android phones
  - Increase in computing power – cloud computing – infinite amount of computing power
  - Storage space is increasing & Storage cost is decreasing
    - More storage you have, the more data you will find to put on it
- Software
  - Web Services (Social media) that allow data collection
    - Big Data
  - Advances in machine learning techniques
  - Availability tools like R + Python



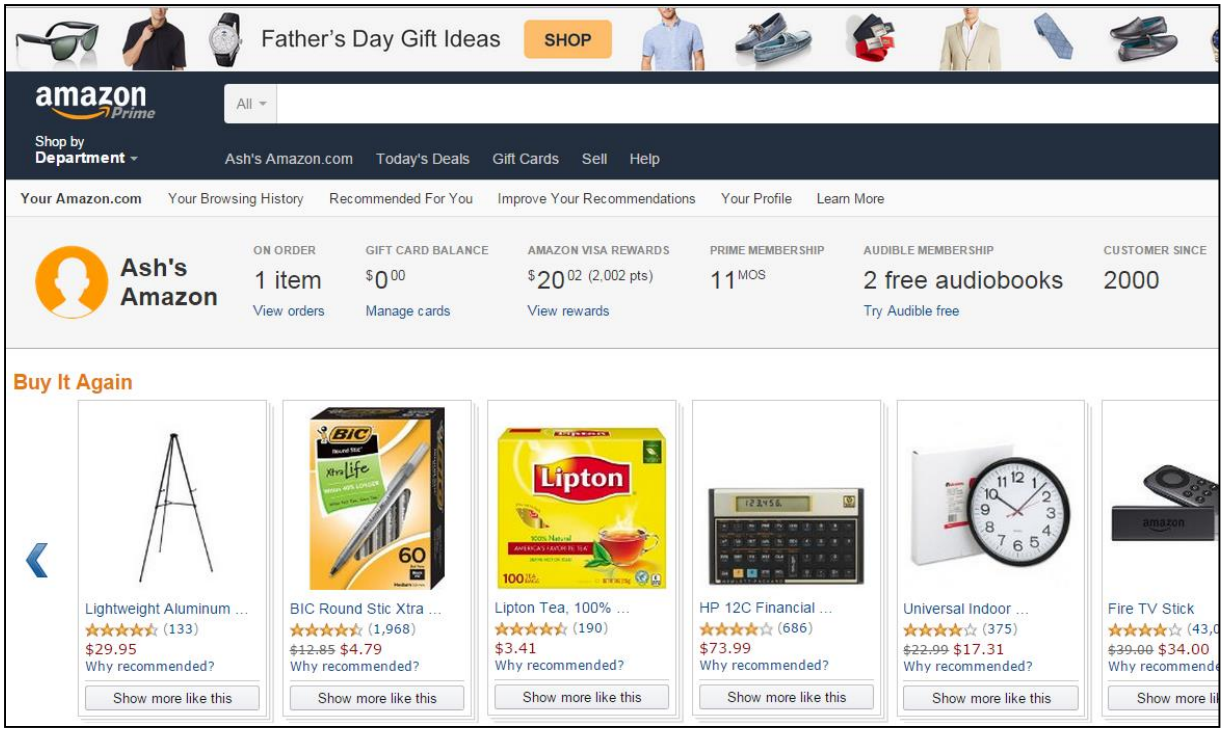
Data Science Applications

Application Areas for Analytics

- Business (customers, products, operations, etc.)
- Manufacturing (Effectiveness and Efficiency)
- Healthcare/Medicine (Clinical, Biological)
- Science (Large Hadron Collider / CERN)
- Entertainment / Sports (Sports Analytics)
- Internet (Social Media, Social Networks)
- Government / National Security
- It is very hard to find an industry that does not have some type of analytics application

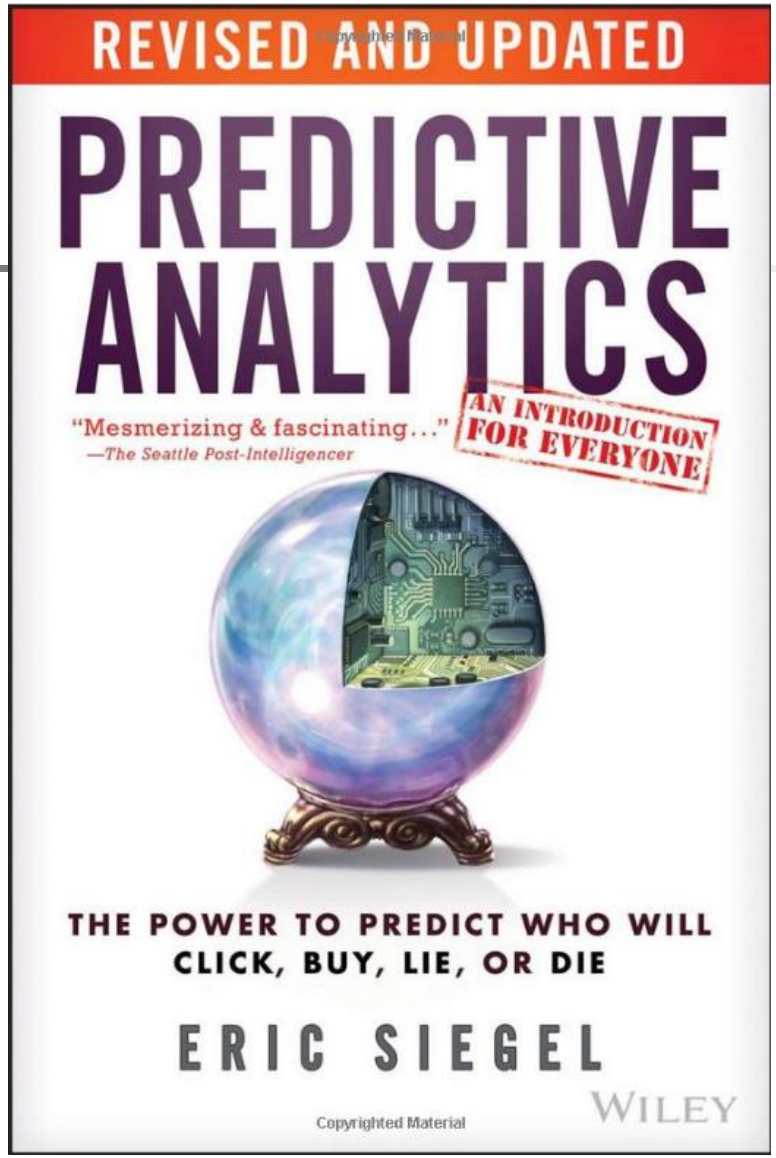
Amazon – Recommender System

- Data
  - Customers past purchase data
- Benefits
  - Recommender system
- Business
  - Higher sales – higher profits



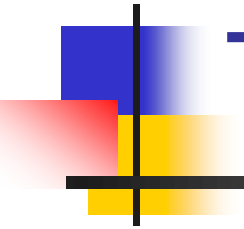
Data Science Predicted Analytics Applications

- Understanding Markets
- Predicting Consumer Choice
- Finding New Customers
- Retaining Customers
- Positioning Products
- Developing New Products
- Promoting New Products
- Recommending Products
- Assessing Brands and Prices
- Utilizing Social Networks
- Watching Competitors

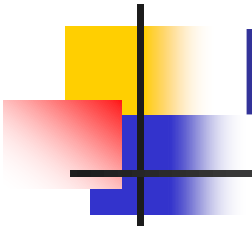

REVISED AND UPDATED
PREDICTIVE ANALYTICS
"Mesmerizing & fascinating..."
—The Seattle Post-Intelligencer
AN INTRODUCTION FOR EVERYONE
THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE
ERIC SIEGEL
WILEY
Copyrighted Material

Applications of Predictive Analytics

1. Family and Personal Life
2. Marketing, Advertising, and the Web
3. Financial Risk and Insurance
4. Healthcare
5. Law Enforcement and Fraud Detection
6. Fault Detection, Safety, and Logistical Efficiency
7. Government, Politics, Nonprofit, and Education
8. Human Language Understanding, Thought, and Psychology
9. Workforce: Staff and Employees



Trends in Technology

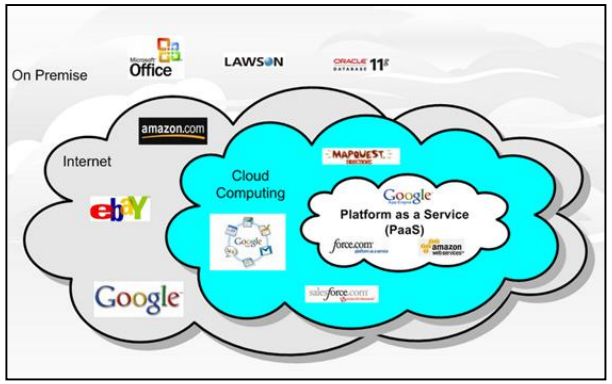


Major Trends

- Internet has made information available to nearly everyone
- Mobile devices can access internet
  - 24/7
- Cloud Computing
  - Infinite computing power

Future Technology Platform

- Mobile devices
  - Laptop / Tablets
  - Smart Phones
- High Speed Cellular Networks
  - 4G / 5G
- Cloud Computing
  - Servers will do all the computing



Hadoop

- Apache Hadoop
  - Open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.
- Hadoop was created by
  - Doug Cutting and Mike Cafarella in 2005.
  - Cutting, who was working at Yahoo! at the time named it after his son's toy elephant.

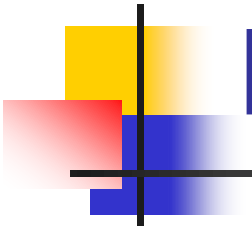
Hadoop: Toddler Talk Provides Big Data Name

Chris Morris, Special to CNBC.com
Tuesday, 28 May 2013 | 11:57 AM ET

CNBC

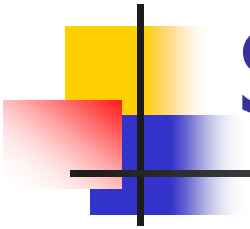
Source: Doug Cutting

Doug Cutting and Hadoop the elephant



Hadoop

- Hadoop changes the economics and the dynamics of large scale computing. Its impact can be boiled down to four salient characteristics.
- Hadoop enables a computing solution that is:
  - Scalable– New nodes can be added as needed
  - Cost effective– Hadoop brings massively parallel computing to commodity servers.
  - Flexible– Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources.
  - Fault tolerant– When you lose a node, the system redirects work to another location of the data and continues processing without missing a fright beat.



Summary

- Big Data and Analytics
- Data Science Applications
- Trends in Technology