Introduction to Data Analytics

Digital Vidya

Agenda

Agenda

Introduction to Data Analytics

- ❖What is Data Analytics
- ❖Evolution of Data Analytics
- ❖Statistics and Data Analytics
- ❖Journey from Data Analyst to Data Scientist
- ❖Industry Use Cases

Terminologies Demystified

- ❖Artificial Intelligence
- ❖Machine Learning
- ❖Deep Learning
- ❖Big Data

Python and Data Analytics

- ❖Introduction
- ❖Python for Data Analytics
- ❖Jupyter Notebook
- ❖Other popular Data Analytics Tools – R , SAS

Introduction to Data Scientists Open Platforms

- ❖What is Kaggle
- ❖Beginner's view into Kaggle
- ❖Introduction to DataScientists.net

Dataset Introduction (30 minutes)

Starting with Python Programming

Introduction to Data Science/Data Analytics

Digital Vidya

Data in today's world

Data is the new Natural Resource

What is Data Analytics?

Set of Processes & Techniques to derive meaningful insights!

What is Analytics

Analysis is the process of breaking a complex topic or substance into smaller parts in order to gain a better understanding of it.

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

Why Businesses need Analytics?

- Humans can count only till so much
- We understand summarized information
- We understand graphs faster
- We need to take decisions
- Wrong Decisions lead to huge costs

Evolution of Data Analytics

Analytics Roles

Statisticians

Business Analysts

Data Analysts

Data Scientists

Big Data Scientists

Technology Landscape

Born in the 1960s

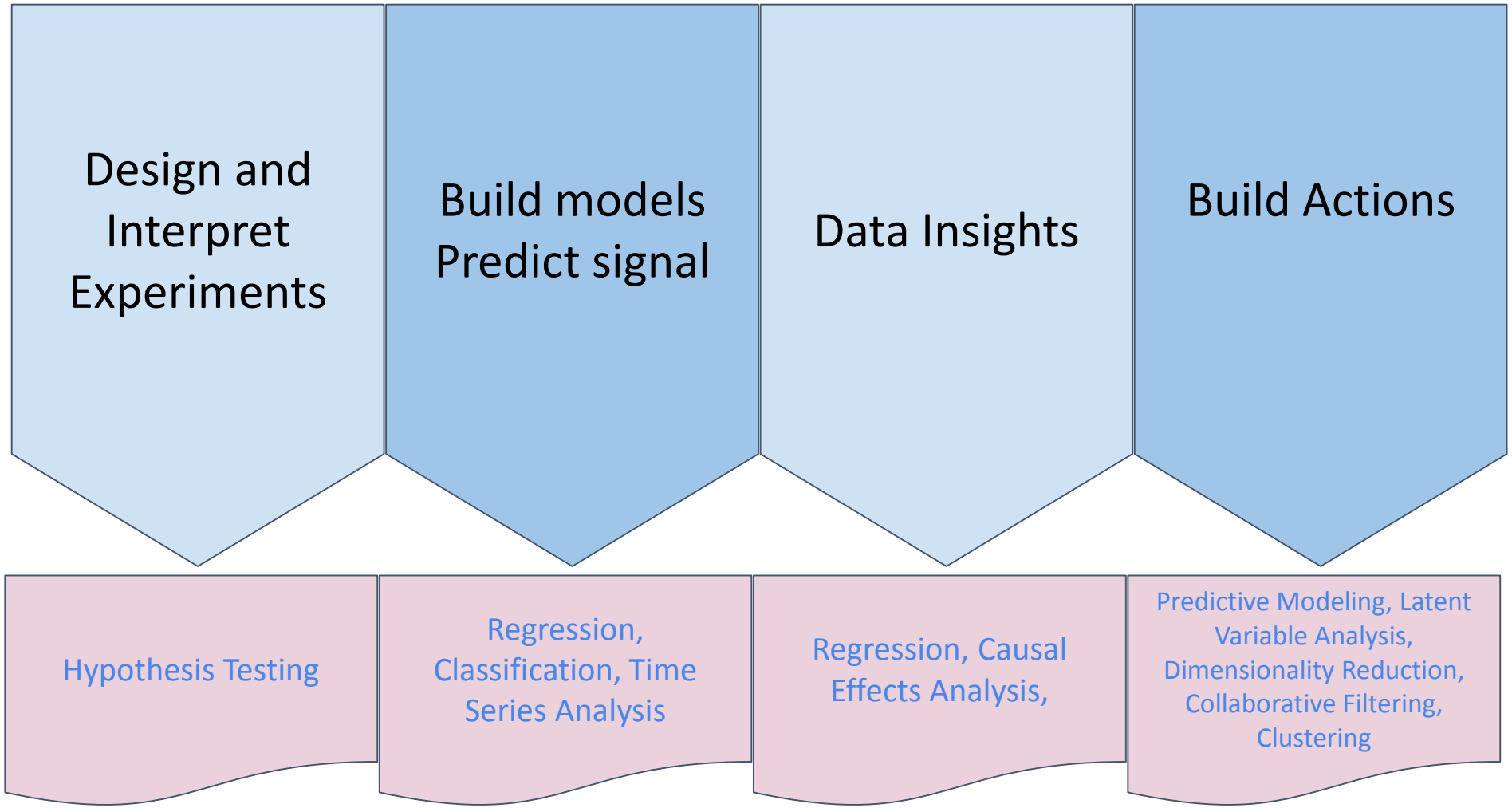
Born in mid 1980s

Born around 1995-2000

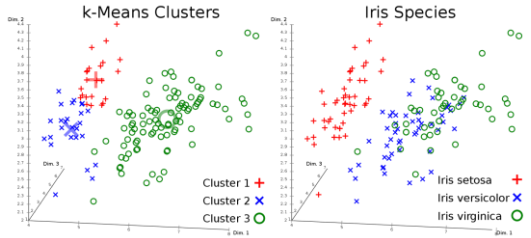
Born around 2009

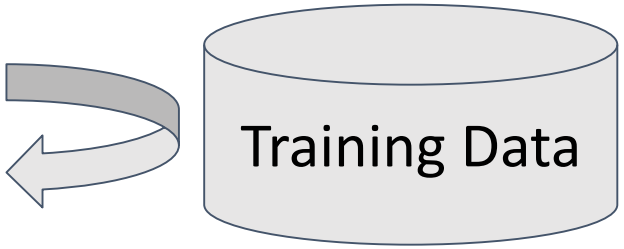
Born around 2010

Statistics and Data Analytics



Machine Learning





Supervised Learning

Unsupervised Learning

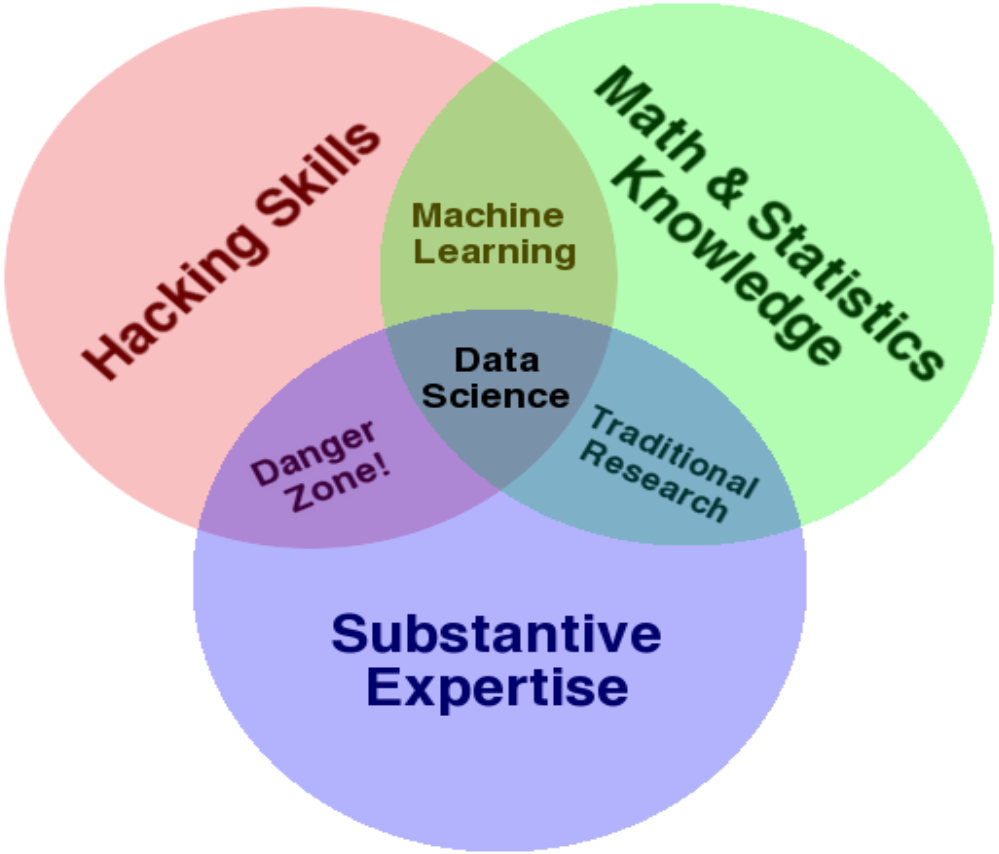
Semi-supervised Learning

Data Analytics Life Cycle

Journey into Data Science

Digital Vidya

Data Science



Hacking (Acquiring, Cleaning Data) + Maths/Statistics (Insights from Data) +
Substantive Expertise (Domain Knowledge, Discovery)

= Data Science

Source: Adopted from the Venn Diagram
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Industry Use Cases

Examples from Industry

Industry	Company	Example
Cab rental	Uber	Recommending drivers where they should place themselves via heatmap in order to take advantage of the best fares and most passengers.
Media and Entertainment	Netflix	Recommender System
Healthcare	Merck	Safe and effective medicines by predicting molecular activity.
Finance	Springleaf	Autonomous offering of personal and auto loans to the customers.
Real estate	Zillow	Home value prediction.
Manufacturing	Bosch	Reduce manufacturing failures.

Terminologies Demystified

Artificial Intelligence

What is Artificial Intelligence (AI) popular definitions:

Theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

Area of computer science that emphasizes the creation of *intelligent* machines that work and react like humans.

Machine Learning

What is Machine Learning (ML) Wikipedia definition:

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. The name machine learning was coined in 1959 by Arthur Samuel.

Machine learning is a way to achieve Artificial Intelligence.

Deep Learning

What is Deep Learning (deeplearning.net):

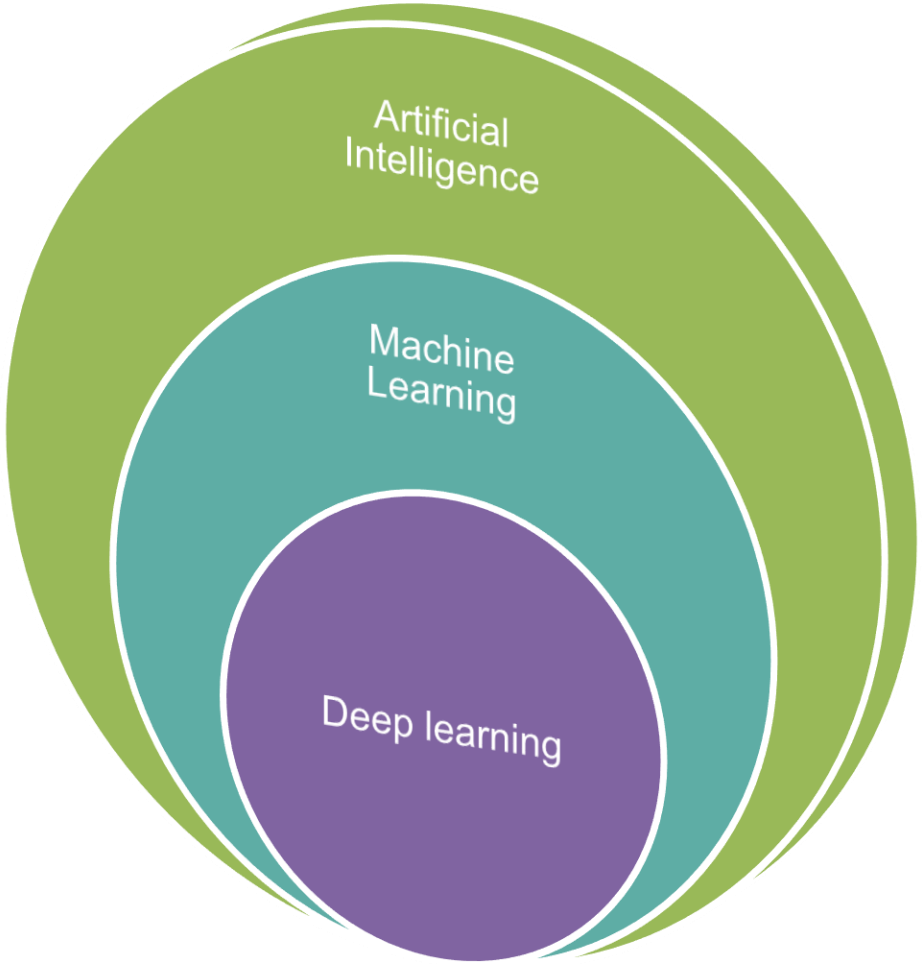
***Deep Learning* is a new area of *Machine Learning* research, which has been introduced with the objective of moving *Machine Learning* closer to one of its original goals: Artificial Intelligence.**

Deep Learning

What is Deep Learning (Oxford University):

“In our quest to build machines capable of different brain functions, such as image and speech understanding, we have discovered that it is of paramount importance to understand how data in the world shapes the brain. Models that are learned from data are the best at many tasks such as image understanding (e.g., knowing where faces occur in images, recognizing road features in self-driving cars) and speech recognition.”

In Summary



Use cases from the world

- **Self driving cars**
  - Tesla autopilot
  - Google self-driving car

- **Virtual Assistants - Siri, Alexa, Cortona, Google Assistant**

- Robots in smart warehouses- Amazon & Alibaba
- Automated Customer Agents & Chatbots
- Image Classification and Recognition Systems

The Era of Big data

Scale of Data

- 40ZB of data will be created by 2020, 300% increase since 2005
- 6 billion people have cell phones
- World population 7 billion

Analysis of streaming data

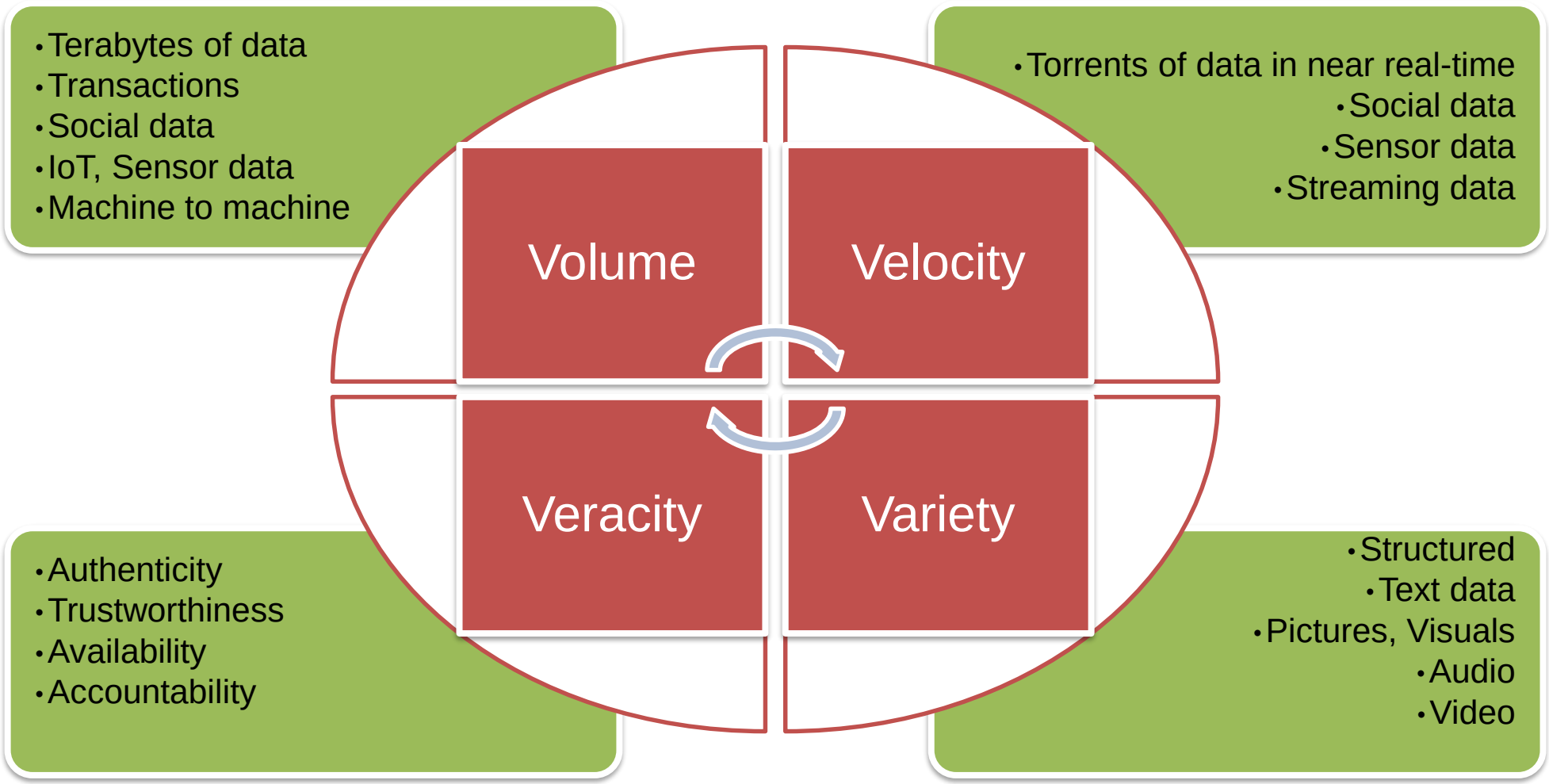
- 1TB of NYSE trade infomation during each trading session
- A jet engine can generate 10TB of data in 30min, 25k flights per day worldwide
- Sensors everywhere - IOT data

Different form of data

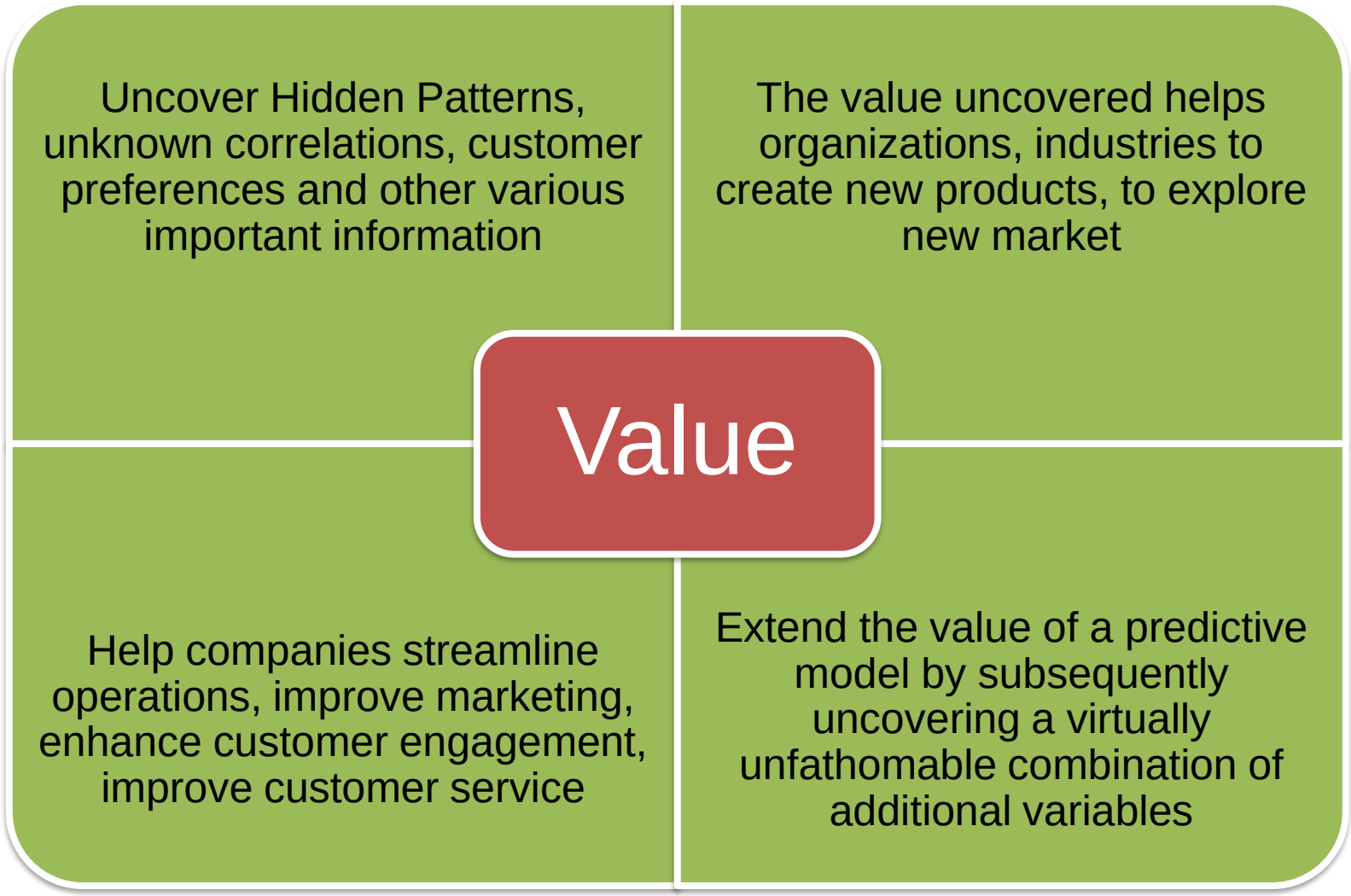
- 150EB of data by 2011 generated by Healthcare domain
- 30 billion content shared on FB every month. 208,300 photos every minute.
- 400 million tweets sent per day by 200 million active users
- 4 billion hours of video is watched on Youtube each month

Digital Vidya

Big Data - The 4 Vs



Big Data - The 5th V



Python and Data Analytics

Digital Vidya

Introduction

Python is a programming language that lets you work quickly and integrate systems more effectively. Used in:

- Statistics
- Data Analysis
- Data Visualization
- Machine Learning
- Deep Learning

Python for Data Analytics

- Statistics i.e statsmodels.
- Mathematics i.e numpy and scipy.
- Data Handling i.e pandas.
- Data Visualization i.e matplotlib, seaborn, plotly and ggplot.
- Machine Learning i.e. Scikit-learn.

Jupyter notebook

- Ipython (Interactive python) Integrated development environment.
- The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser.
- The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.

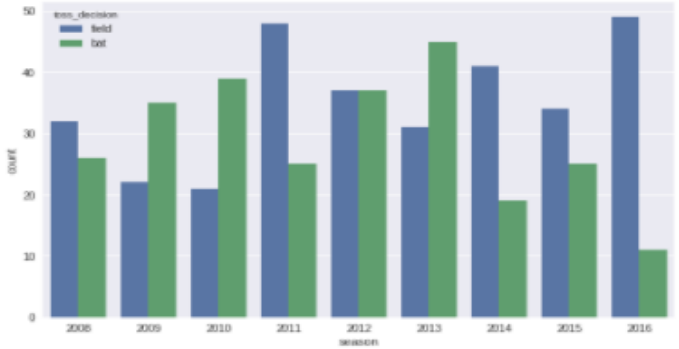
Toss Decisions

In [11]:
```
print('Toss Decisions in %\n',((matches['toss_decision']).value_counts())/577*100)
```

```
Toss Decisions in %
 field    54.592721
bat       45.407279
Name: toss_decision, dtype: float64
```

Toss Decisions across Seasons

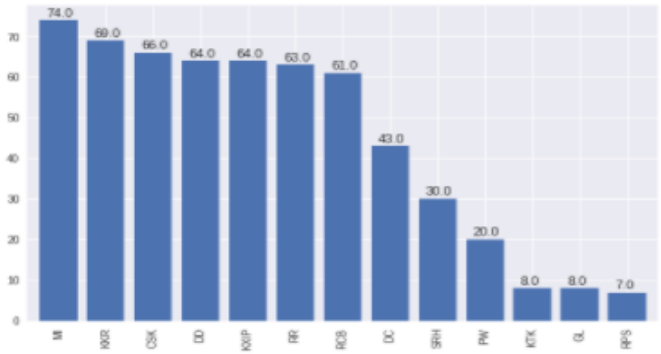
In [12]:
```
mlt.subplots(figsize=(10,6))
sns.countplot(x='season',hue='toss_decision',data=matches)
mlt.show()
```



The decision for batting or fielding varies largely across the seasons. In some seasons, the probablity that toss winners opt for batting is high, while it is not the case in other seasons. In 2016 though, the majority of toss winners opted for batting.

Maximum Toss Winners

In [13]:
```
mlt.subplots(figsize=(10,6))
ax=matches['toss_winner'].value_counts().plot.bar(width=0.8)
for p in ax.patches:
    ax.annotate(format(p.get_height()), (p.get_x()+0.15, p.get_height()+1))
mlt.show()
```



Mumbai Indians seem to be very lucky having the higest win in tosses follwed by Kolkata Knight Riders. Pune Supergiants have the lowest wins as they have played the lowest matches also. This does not show the higher chances of winning the toss as the number of matches played by each team is uneven.

Text Markdown Cell

Code Cell

Output Cell

jupyter

Digital Vidya

Other popular languages

- **R** is an open source programming language and software environment for statistical computing and visualization.

- **SAS** language for the Statistical Analysis System is a fourth-generation proprietary programming language

Introduction to Open Platforms for Data Scientists

Digital Vidya

Establishing yourself as a Data Scientist

BEFORE

- STEP 01 — Starting to learn a new skill/language
- STEP 02 — Do coding, and assignments and project
- STEP 03 — Prepare Interview Questions
- STEP 04 — Start with your job process

NOW

- STEP 01 — Starting to learn a new skill/language
- STEP 02 — Do coding, & assignments & project
- STEP 03 — Establish yourself as a coder on platforms like stackoverflow
- STEP 04 — Practice on datasets from different domains on Kaggle like platforms

  Contribute to Kaggle with datasets and kernels (code)

  Participate in Code Competitions

Established Data Scientist

What is Kaggle

- A platform for Data Scientists and Data Science Competitions
- In 2010, **Kaggle** was founded as a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models (Wikipedia)

kaggle

Digital Vidya

Kaggle terminologies

- Datasets
  - ❖ Real world data for open source community
  - ❖ Every Dataset needs a story that catches interest of Data Scientists:
    - Why is this dataset interesting?
    - What questions can the community help with?
  - ❖ 650 + Datasets

kaggle

Digital Vidya

Kaggle Popular datasets

- Some popular datasets
  - IMDB Movie dataset (Entertainment)
  - European Soccer Database (Sports)
  - Credit Card Fraud Detection (Finance)
  - Human Resources Analytics (Cross Industry)
  - Iris (Botany)
  - Climate Change
  - World University Rankings
  - Medical Appointments No shows (Healthcare)

kaggle

Kaggle terminologies

- Kaggle Kernels
  - ❖Code in R or Python
  - ❖Kernels is preloaded with the most common data science languages and libraries.
  - ❖Look at Kernels from peer Data Scientists
  - ❖Look at the Kernels that have most votes
  - ❖Fork from an existing kernel

kaggle

Kaggle Competitions

- Prize money
- Real world Data given by companies that are looking for some serious insights and problem solving
- Examples
  - ❖ Zillow Price (Real-Estate)
  - ❖ Instacart Market Analysis (eCommerce)
  - ❖ Mercedes-Benz (Manufacturing)
  - ❖ Intel and MobileODT Cervical Cancer Screening (Healthcare)

kaggle

Walkthrough of an example

Visit https://www.kaggle.com/c/titanic

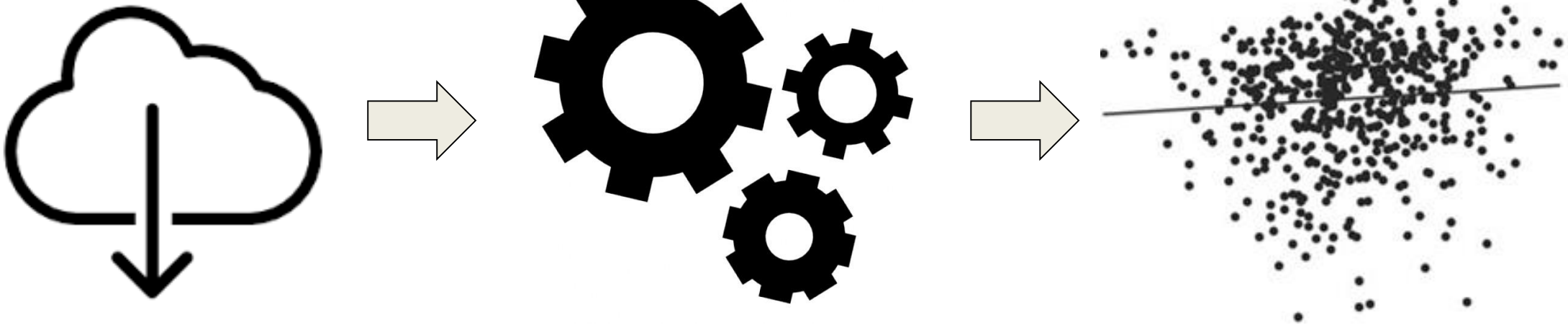
kaggle

Digital Vidya

Datascience.net

- Smaller scale competitions.
- Good for amatuer data scientists/analysts.
- Visit https://www.datascience.net/

How it works?



Download Data

Build Model

Upload and
Test Results

Dataset Introduction for our course

Dataset

- Multiple datasets used to teach several concepts
- You can start to get familiar with this one
  - Ball by ball dataset of IPL until season 9.
  - Download at https://www.kaggle.com/manasgarg/ipl

Motivation!

Predicting the winner of next IPL season based on previous seasons data.

It has been done before!

Google predicted Germany as the winner of FIFA 2014 world cup.
https://googleblog.blogspot.fr/2014/07/google-cloud-platform-predicts-world.html