Data Analytics/Science

Data Sources: Transactional Data

	7	NODE_ID	CATEGORY	SUBCATEGORY	TYPE	FRAUDULENT_TRANS_NUM	TOTAL_TRANS_NUM	GROUP
75	■	Customer075	Customer	25to34	c	3	15	4
76	■	Customer076	Customer	25to34	c	1	16	4
77	■	Customer077	Customer	Under25	c	2	11	4
78	■	Customer078	Customer	45to54	c	4	17	4
79	■	Customer079	Customer	65andOver	c	2	9	4
80	■	Customer080	Customer	25to34	c	4	13	4
81	■	Merchant0001	Retail	DrugStores	m	0	1	1
82	■	Merchant0002	Retail	FoodStore	m	0	1	1
83	■	Merchant0003	Services	Restaurants	m	1	1	1
84	■	Merchant0004	Services	Restaurants	m	0	1	1
85	■	Merchant0005	Services	OtherServices	m	0	1	1
86	■	Merchant0006	Services	OtherServices	m	0	1	1
87	■	Merchant0007	Retail	General	m	1	1	1
88	■	Merchant0008	Services	OtherServices	m	0	1	1
89	■	Merchant0009	Retail	GasStation	m	1	1	1

- The products purchased
- The customers and items details

Data Sources: Social networks

- Posts in social media
- Pictures and videos posted online
- Instant & email messages
- Voice data

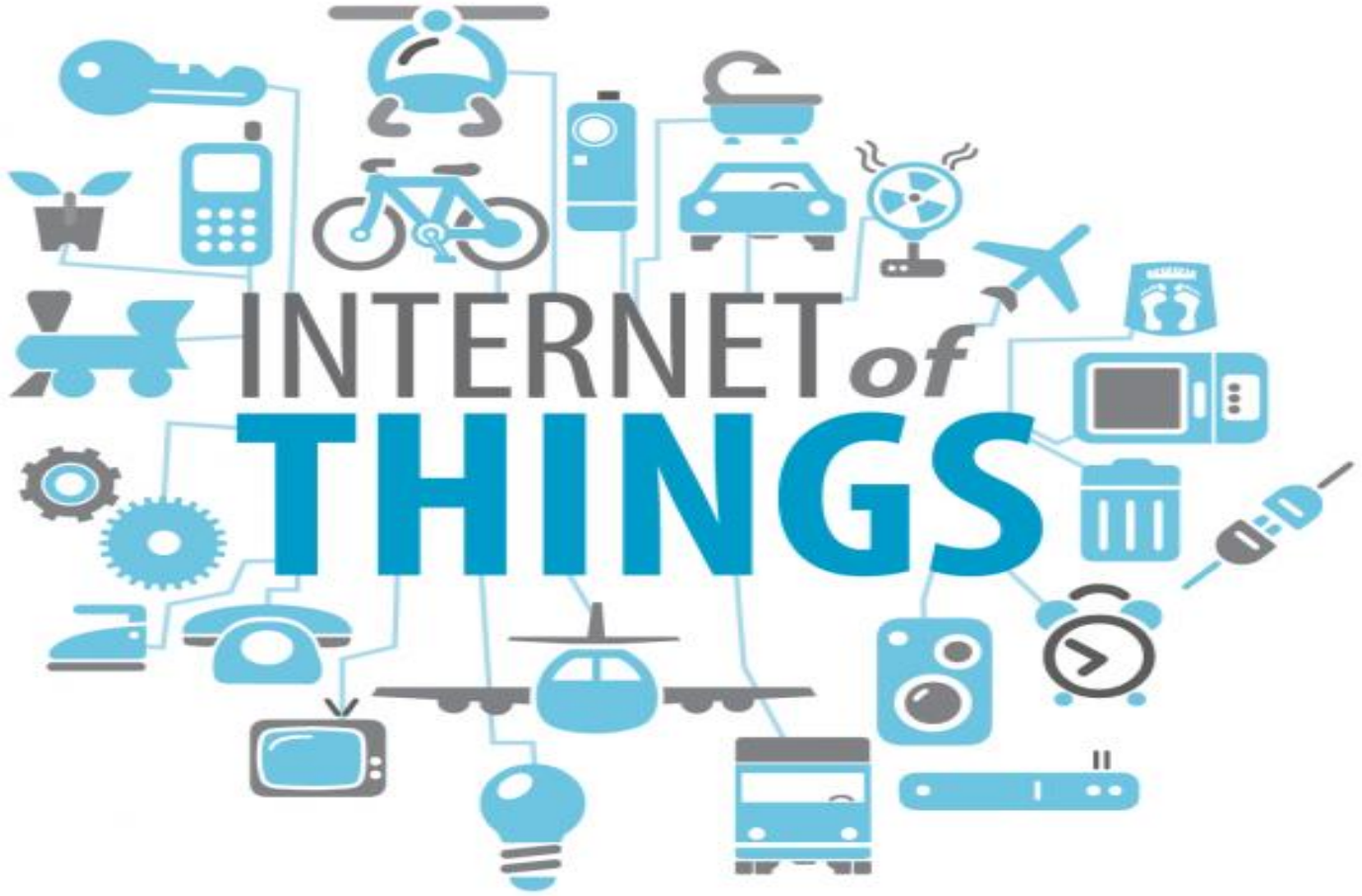
Data Sources: Mobiles

- GPS data
- Mobile apps data
- Voice data

Data Sources: Server logs

```
9/1/99, 10:46:11, 1578, 509, 5397, 200, 0, GET, /cfdocs/akonline/paintbrush.JPG, -,
9/1/99, 10:46:49, 37703, 577, 24402, 200, 0, GET, /cfdocs/akonline/email_book.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:49:11, 181500, 579, 114331, 200, 0, GET, /cfdocs/akonline/update_table.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:52:04, 354641, 662, 163301, 200, 64, GET, /cfdocs/AKONLINE/assess/PAT11B.pdf
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:52:31, 20921, 609, 163301, 200, 0, GET, /cfdocs/AKONLINE/assess/PAT11B.pdf,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:55:30, 178985, 658, 4314, 200, 0, GET, /cfdocs/akonline/adobe_get.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:55:58, 8437, 583, 39430, 200, 0, GET, /cfdocs/akonline/image.JPG, -,
9/1/99, 11:30:25, 5422, 662, 172695, 200, 0, GET, /cfdocs/AKONLINE/assess/TBT11B.pdf,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 11:33:17, 171359, 437, 172695, 200, 0, GET, /cfdocs/AKONLINE/assess/TBT11B.pdf,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 11:40:46, 449531, 582, 1441, 200, 0, GET, /cfdocs/akonline/chooseTableMenu.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 11:41:16, 812, 775, 438, 200, 0, POST, /cfdocs/akonline/exampleTable.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis
```

Data Sources: IoT



More than 65 billion devices were connected to the Internet by 2010, and this number will go up to 230 billion by 2020

What kind of data generated by these sources?

- Structured: Transactional data
- Semi-Structured : Log data, XML & JSON data, Sensor data
- Unstructured: Images, Voice, Video & Text data(chats, emails, Blogs, etc.,)

Why do we care about data?

Why data?

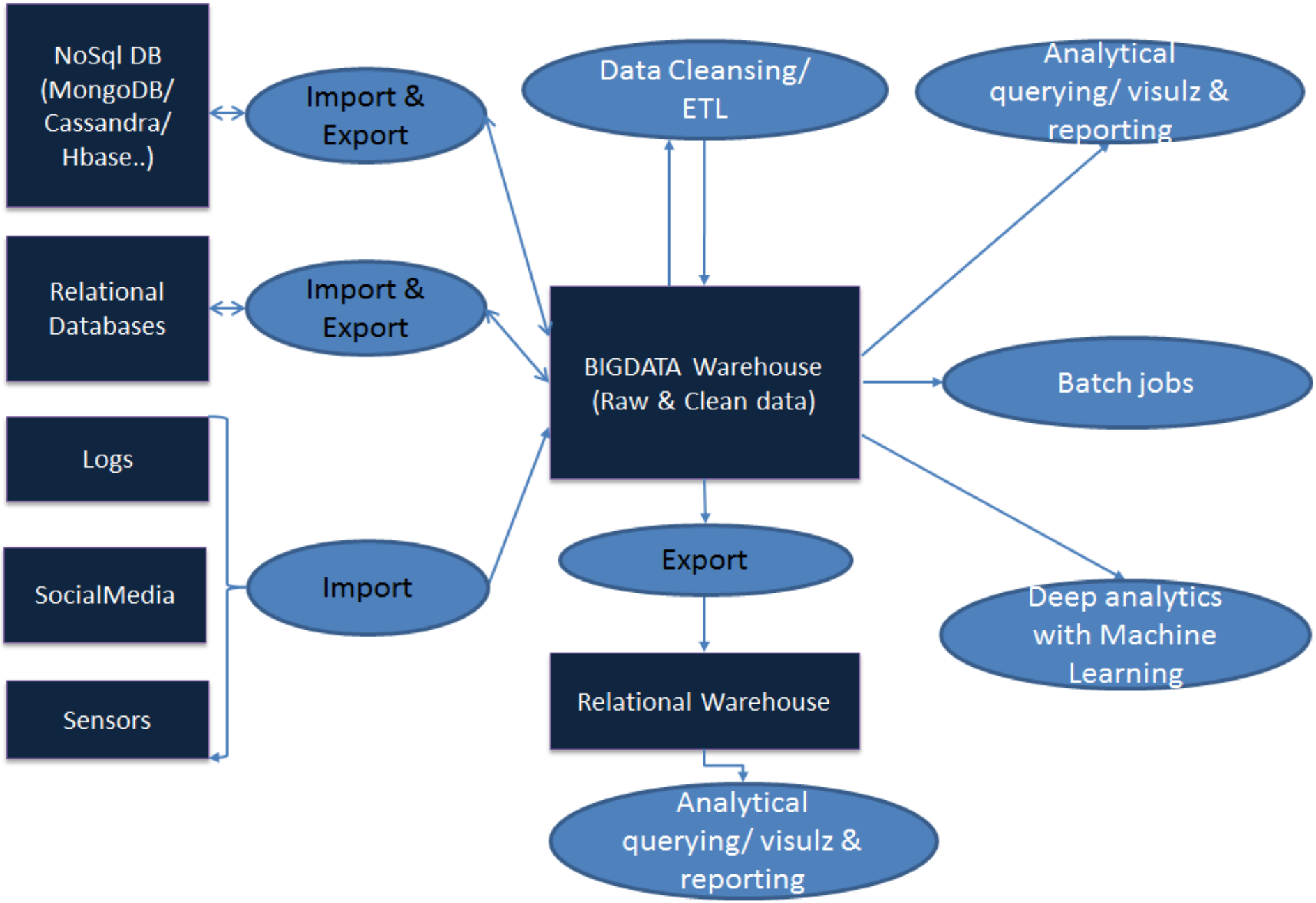
- Data has inherent value and cannot be discarded
  - Get insights from data to offer a better product
  - Get insights from data to make better decisions
  - Take a competitive advantage by providing personalized services

How do you derive value?

- The team of data analysts/scientists does data analysis to derive inherent value in data.

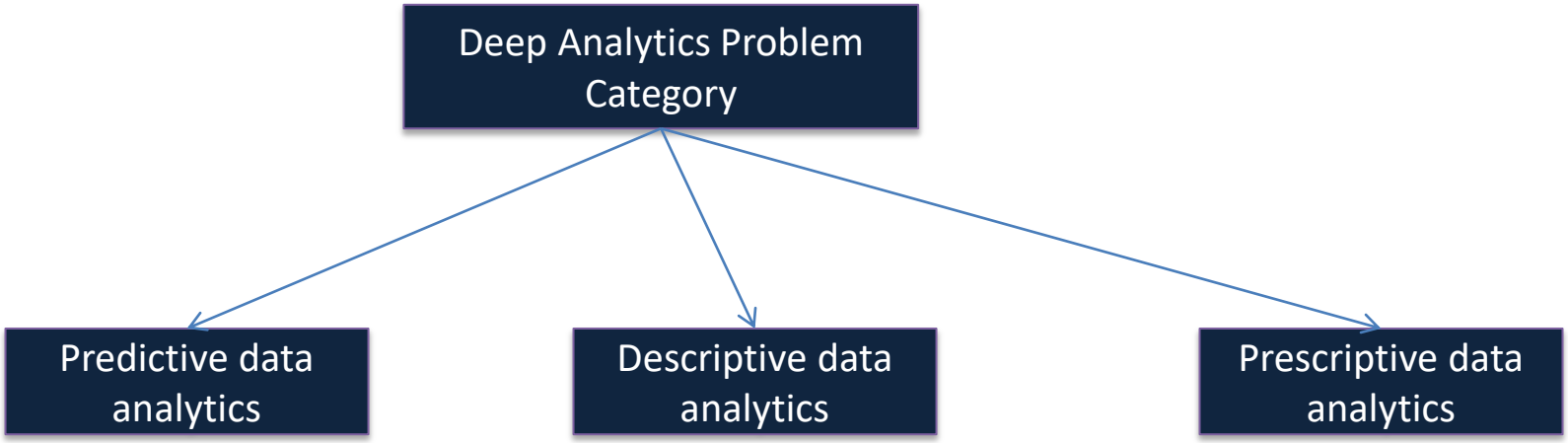
- Does humans or machines derive the insights?

(BIG)Data Analytics: Big Picture

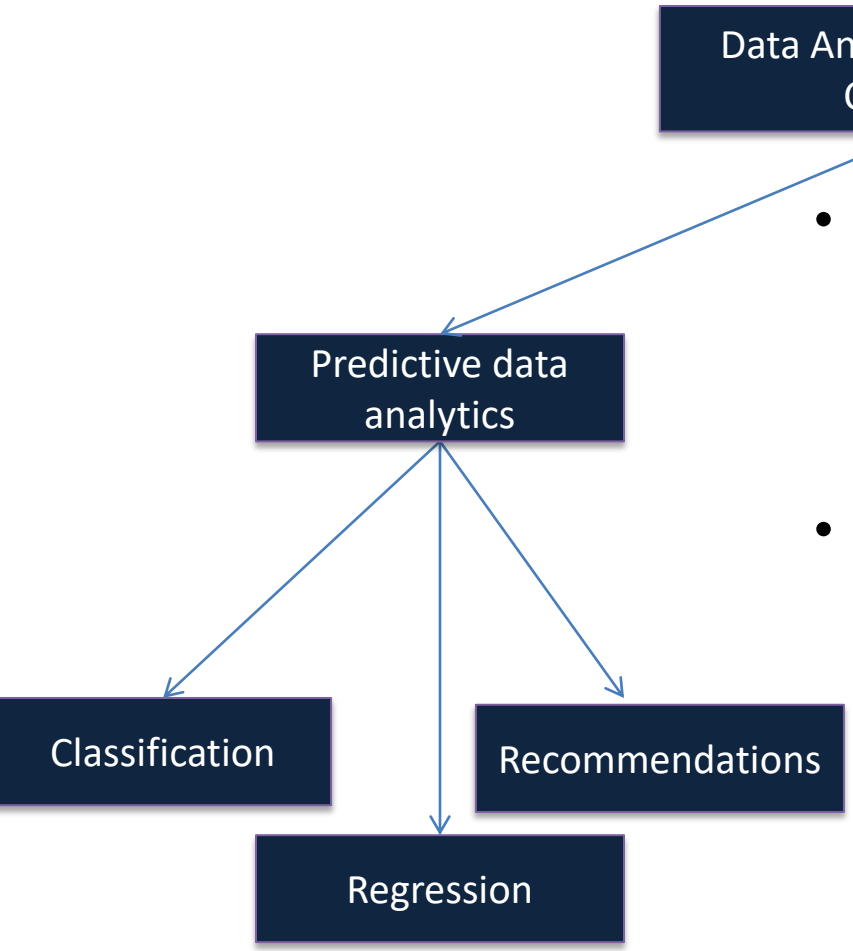


Deep data analytics use-cases

Deep data analytics: overview



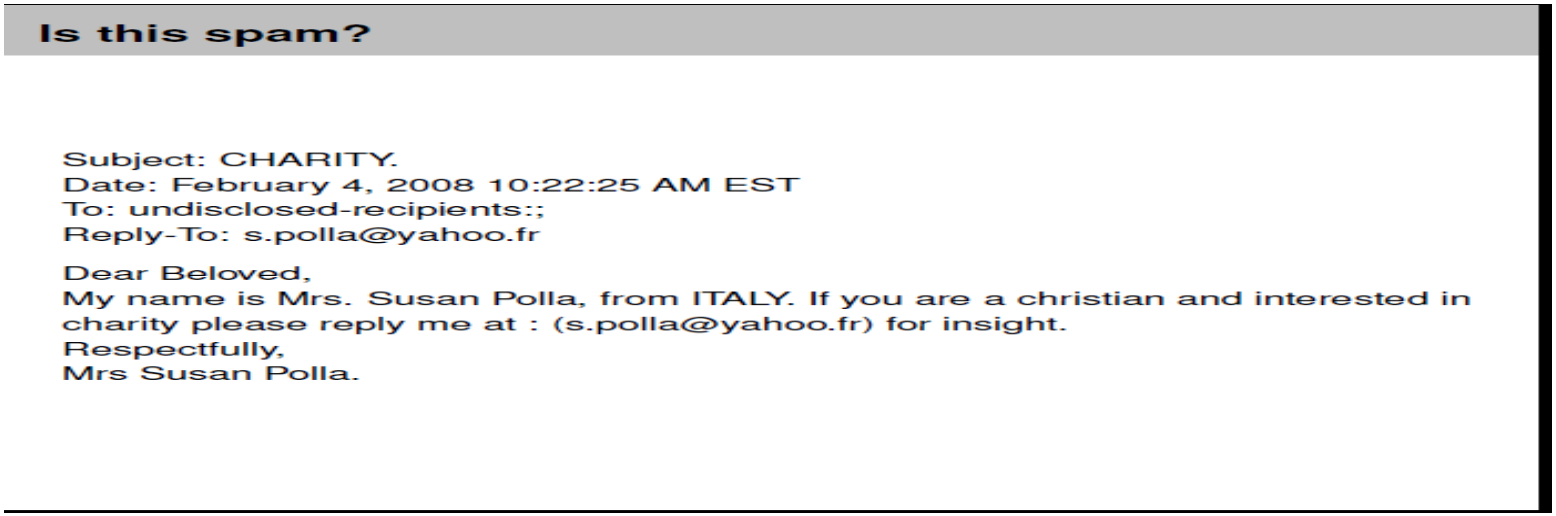
Predictive data analytics



- Predictive analytics uses history data to determine the probable future outcome of an event or chance of a situation occurring
- Predictive analytics focus on predicting a single customer/event behavior

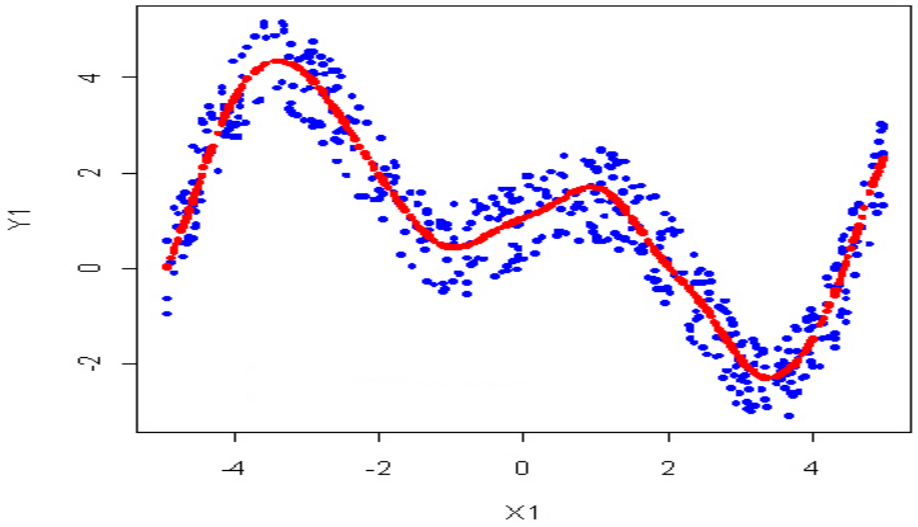
Predictive analytics: Classification

- Classification: Predict the category of an unlabeled observation by analyzing the history of labeled observations
  - Predicting whether new email is spam or not based on past labeled email history
  - Predicting whether customer will be defaulter or not based on past labeled customer history
  - Classifying an image is animal or not based on past labeled image history



Predictive analytics: Regression

- Regression: Predict the numerical value of an unlabeled observation by analyzing the history of labeled observations
  - Predicting the unknown stock price at any time based on history of labeled stock prices
  - Predicting the rating of a non-rated product/movie based on history of labeled products/movies
  - Predicting the value of real-estate property in future based on history of labeled real-estate properties



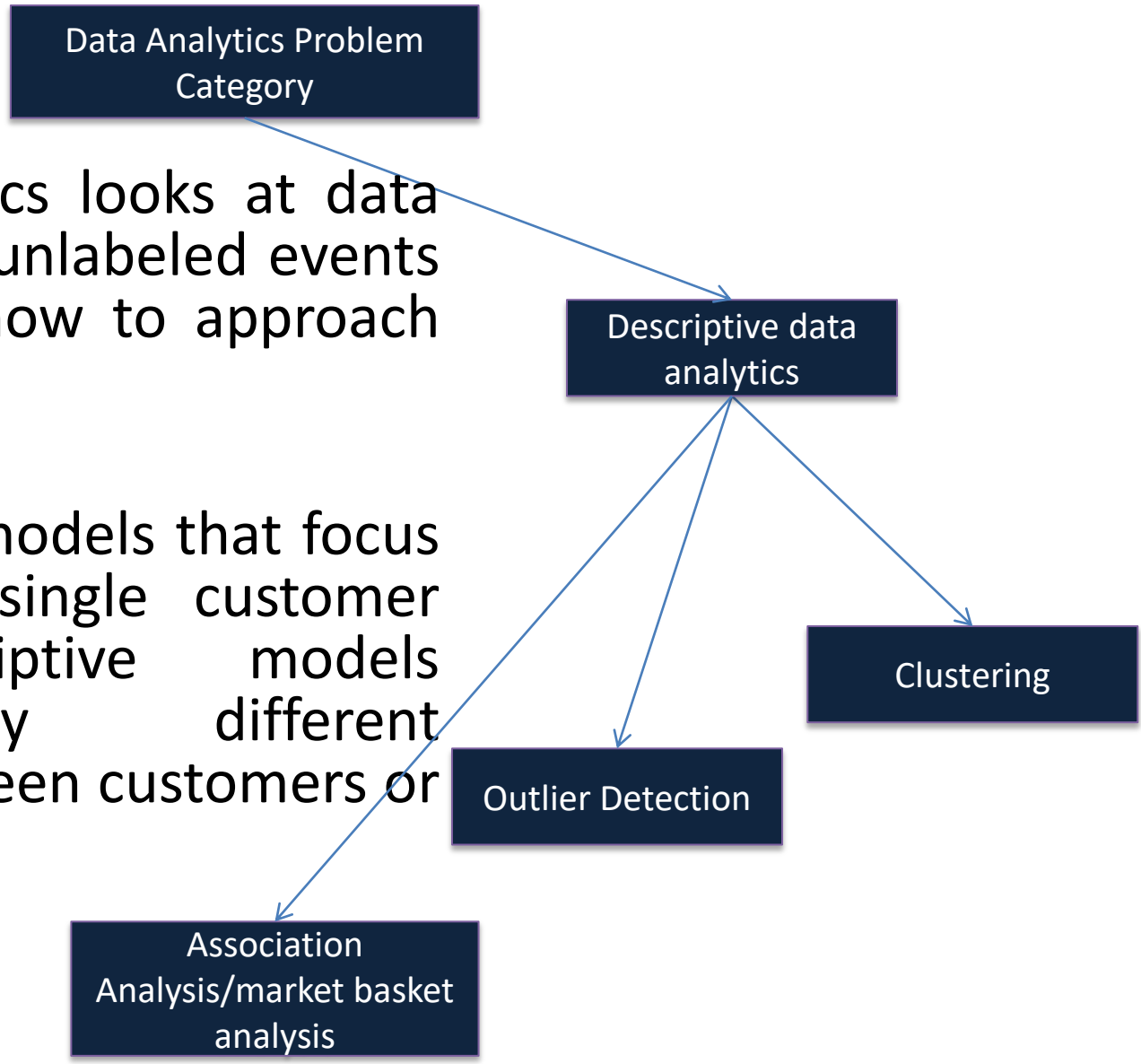
Predictive analytics: Recommenders

- Recommenders: For an user, suggest a bunch of un-purchased items based on history of purchased items by that user
  - Recommend set of unwatched movies based on history of movies watched by that user

Descriptive data analytics

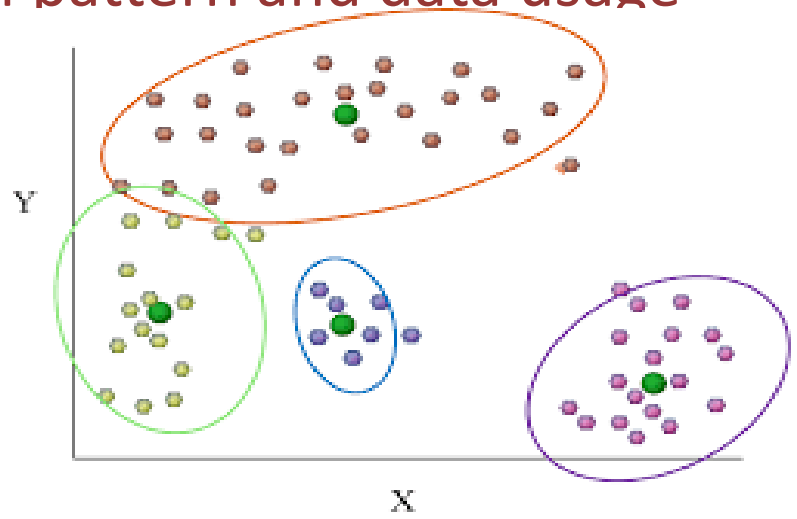
- Descriptive analytics looks at data and analyzes past unlabeled events for insight as to how to approach the future.

- Unlike predictive models that focus on predicting a single customer behavior, descriptive models identify many different relationships between customers or products.



Descriptive analytics: Clustering

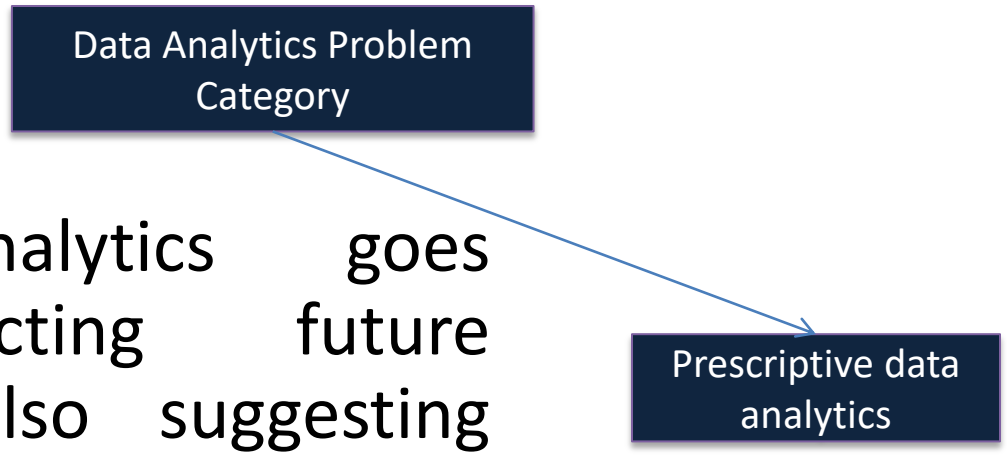
- Clustering: Find the groups of related events from the history of unlabeled events
  - Find the different groups of document clusters from unlabeled document collections
  - Find the groups of similar search results from the entire unlabeled search results
  - Find the different groups of telecom subscribers based on their call pattern and data usage



Predictive analytics: Associations

- Association Analysis: For a customer, recommend the items that are frequently bought together with the current item by analyzing the transactions of customers who purchased that item
  - Recommend the movies watched together with the one you are viewing/searching now based on past viewing history of that movie
  - Suggest the items to place together in a store based on items purchased together by customers

Prescriptive data analytics

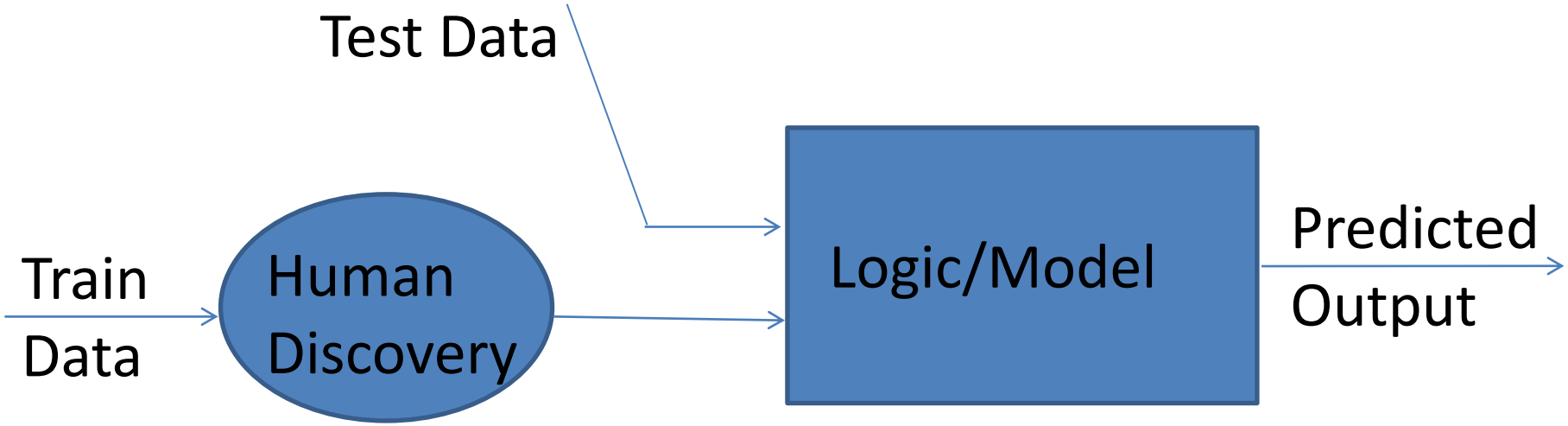


- Prescriptive analytics goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the decision maker the implications of each decision option.

How do you solve data analytics problems?

- . . . Lets start our journey with predictive analytics problems first

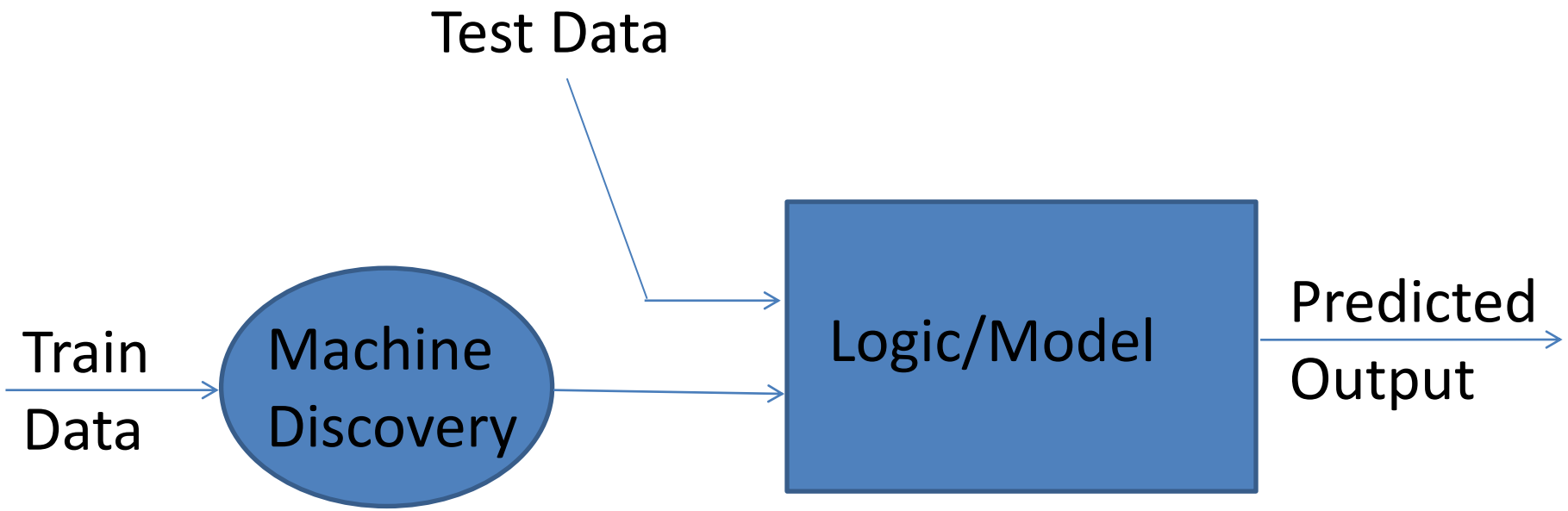
Traditional Approach: Human discovery



Issues:

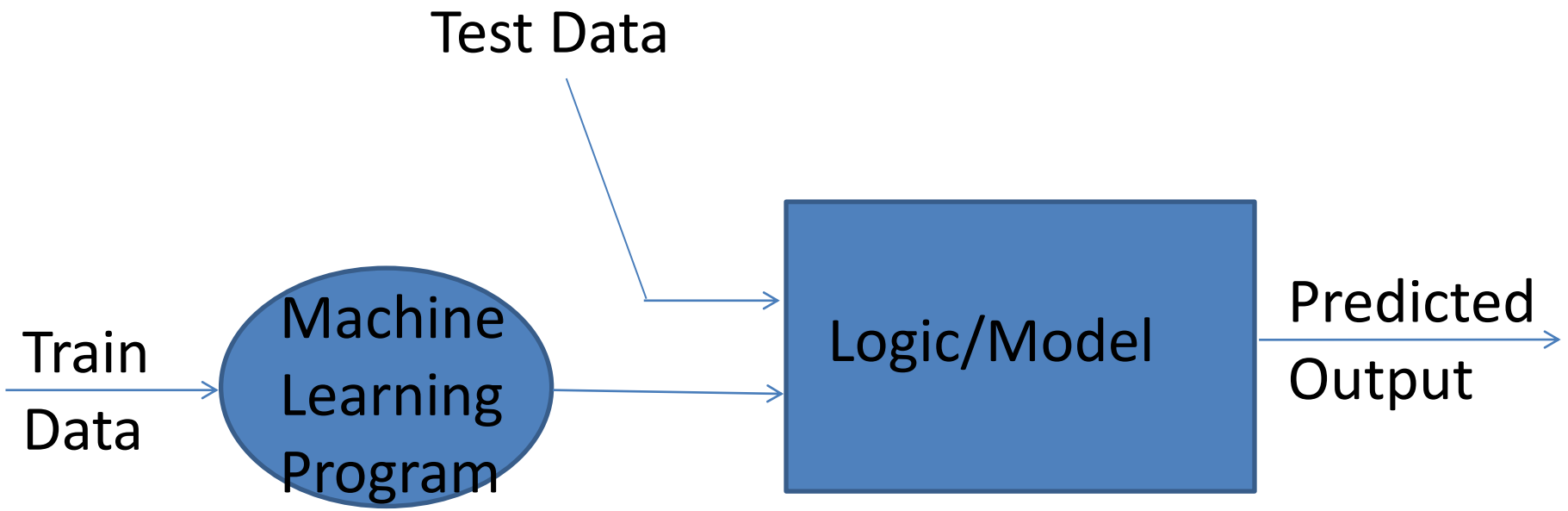
- Human has to manually go through the train data to discover pattern or logic that can be applicable for test data.
- Hard-coded logic and needs to rediscover and update the logic whenever train data gets changed.
- Is it practical to discover the logic/pattern if train data is big?

New Approach: Machine discovery



- If machine could discover logic/pattern, its great!!
  - No hardcoded logic
  - Machine can rediscover logic if train data gets changed
  - Machines can process big data sets(with more peers) as well.
- But How does machine discovers logic/pattern?

New Approach: Machine discovery



- Any thing in machine happens via software. So, we have to write programs that discovers logic automatically. We call it as Machine learning programming.

- Different Machine learning approaches discover logic/pattern in different ways. We have to find what approach is best for discovering logic in given train data.

Data Analytics Lifecycle